Linear Systems

Professor Yi Ma
Professor Claire Tomlin
GSI: Somil Bansal
Scribe: Chih-Yuan Chiu

Department of Electrical Engineering and Computer Sciences
University of California, Berkeley
Berkeley, CA, U.S.A.

May 22, 2019

Contents

Preface

The Fall 2018 graduate-level EE221A Linear Systems Theory course, offered by Professor Yi Ma in the Department of Electrical Engineering and Computer Sciences (EECS) at the University of California, Berkeley, included content largely drawn from the following sources:

- Tomlin, Claire. *Lecture Notes on Linear Systems Theory* [10].
- Ma, Yi. *Lectures Notes on Linear System Theory* [7].
- Callier, Frank and Desoer, Charles. *Linear System Theory* [4].
- Sastry, Shankar. *Nonlinear Systems: Stability, Analysis, and Control* [9]
- Liberzon, Daniel. *Calculus of Variations and Optimal Control, A Concise Introduction* [6]
- Yung, Chee-Fai. *Linear Algebra, 3rd Edition.* [11]
- Yung, Chee-Fai. *Lecture Notes on Mathematical Control Theory.* [12]

The following collection of notes represents my attempt to organize this ensemble of linear systems-related material into a friendly introduction to the subject of linear systems. The chapters contain collections of lectures in Professor Claire Tomlin's *Lecture Notes on Linear Systems Theory* [10], presented in roughly the same order, with the exception of Lecture 11 (Linear Quadratic Regulator), which has been relocated to the end of the text. Chapter 1 contains content from Lecture 1, which gives an introduction to linear systems. Chapter 2 includes material from Lectures 2-6, and primarily reviews concepts in linear algebra, such as fields, vector spaces, linear independence and dependence, linear maps and matrix representations, norms, orthogonality, inner product spaces, adjoint maps, projection, least squares optimization, and the singular value decomposition. Chapter 3 was organized from Lectures 7-10, and formally introduces dynamical systems and their properties, beginning with the Fundamental Theorem of Differential Equations, particular classes of dynamical systems (linear, non-linear, time-invariant, time-varying), and concluding with properties of the matrix exponential and an inverted pendulum example. Chapter 4, compiled from material in Lectures 12-15, discusses notions of stability, as well as necessary and sufficient conditions for these different definitions of stability. Chapter 5, which spans Lectures 16-22, defines controllability, observability, stabilizability, and detectability, and explores different criteria that different types of systems

must satisfy in order to be controllable, observable, stabilizable, and/or detectable. Chapter 6 discusses the Linear Quadratic Regulator, the subject of Lecture 11, as well as the Hamilton-Jacobi-Bellman Equation, as discussed in Chapter 2 of [6]. Finally, appendices including the Basic Lypaunov Theorem and other stability theorems for non-linear systems, among other material, have been added, partly for completeness, and partly to interest readers in more advanced topics in control theory. This material originates largely from Chapter 5 of [9].

These notes have several possible shortcomings. To minimize the reader's confusion, I have attempted to unify the notation used in the references cited above, and correct most of the typos in the text. Nevertheless, it is inevitable that minor errors or inconsistencies in notation remain scattered throughout the notes (Readers have discovered such mistakes are welcome to contact me at chihyuan_chiu@berkeley.edu). Regarding the material itself, Professors Yi Ma and Claire Tomlin often gave useful remarks in their lectures that were not included in their written notes. Although I have added as many of these comments into the notes as possible, it is certain that I have missed many others. Moreover, some supplementary material from the Fall 2017 EE221 Linear Systems Theory course were omitted, since I found the material to be similar to content already included. (e.g. Somil Bansal's "Special Lecture on the Linear Quadratic Regulator," which discusses dynamic programming solutions to the finite LQR problem, and Dr. Jerry Ding's "Alternative Derivation of Linear Quadratic Regulator," which describes how the Pontryagin Minimum Principle can be applied to solve the linear quadratic optimization problem.) My inexperience with the subject of linear systems may also have contributed towards errors in the notes. Nonetheless, it is my hope that the text remains a useful, introductory reference to readers studying linear systems theory for the first time.

It is an honor for me to dedicate these notes to the following individuals. Naturally, without the carefully prepared lectures and handouts given by Professor Yi Ma, and the painstakingly detailed notes and figures organized by Professor Claire Tomlin, this work would not exist. I would also like to acknowledge Professors Shankar Sastry, author of "Nonlinear Systems: Stability, Analysis, and Control" [9], and Professor Daniel Liberzon, author of "Calculus of Variations and Optimal Control, A Concise Introduction" [6], for their time and effort into compiling these works, which serve as the foundation of most of the material in the last chapter and appendix. Somil Bansal, the Graduate Student Instructor for this course in the Fall semester of 2018, deserves gratitude not only for meticulously preparing the discussion notes included in this text, but also for taking the extra effort to arrange office hours, and organize midterm and final discussion sessions. My appreciation extends to Professor Chee-Fai Yung, who first sparked my interest in control theory, and who graciously allowed me to use sections of his lectures notes in this work. I would also like to acknowledge my fellow classmates in this course, many of whom provided helpful suggestions throughout the semester. Lastly, I would like to thank my parents for their unending support and encouragement.

Chih-Yuan Chiu
University of California, Berkeley
Department of Electrical Engineering and Computer Sciences
December 2018

Notation

The following notation will be employed throughout this text:

$\mathbb{N}$: Set of all positive integers
$\mathbb{Z}$: Set of all integers'
$\mathbb{Q}$: Set of all rational numbers
$\mathbb{R}$: Set of all real numbers, i.e. the real line
$\mathbb{C}$: Set of all complex numbers
$\mathbb{C}^-$: Set of all complex numbers with a (strictly) negative real part
$\overline{\mathbb{C}^-}$: Set of all complex numbers with a non-positive real part
$\mathbb{C}^0$: Set of all purely imaginary numbers
$\mathbb{C}^+$: Set of all complex numbers with a (strictly) positive real part
$\overline{\mathbb{C}^+}$: Set of all complex numbers with a non-negative real part
$\in$: Is an element of (Is contained in)
$\forall$: For each (for all)
$\exists$: There exists
$\exists!$: There exists a unique
$\exists?$: Does there exist
$\ni$: Such that
$A \Rightarrow B$: A implies B
$B \Leftarrow A$: B implies A
$A \Leftrightarrow B$: A and B are equivalent
I_n: Identity matrix of dimension $n \times n$
O_n: Zero matrix of dimension $n \times n$
$\circ$: Composition of Functions
$1_{\mathcal{X}}$: Identity map from $\mathcal{X}$ to $\mathcal{X}$
$S_1 \subset S_2$: The set S_1 is a subset of the set S_2
$\mathcal{W} \leq \mathcal{V}$: The vector space $\mathcal{W}$ is a subspace of the vector space $\mathcal{V}$
$\mathcal{W} \oplus \mathcal{V}$: The direct sum of $\mathcal{W}$ and $\mathcal{V}$.
$\mathcal{W} \overset{\perp}{\oplus} \mathcal{V}$: The orthogonal direct sum of $\mathcal{W}$ and $\mathcal{V}$.
$|\cdot|$: Norm of a vector
$\|\cdot\|$: Norm of a matrix or operator
$X(s)$: Unilateral Laplace transform of $x(t)$
(If the time-domain argument is capitalized, e.g. $X(t)$, a hat is used, e.g. $\hat{X}(s)$).
$u_{st}(t)$: Unit step function

LTI: Linear time-invariant
LTV: Linear time-variant
SISO: Single-input-single-output
MIMO: Multiple-input-multiple-output

Chapter 1

Introduction

1.1 Lecture 1

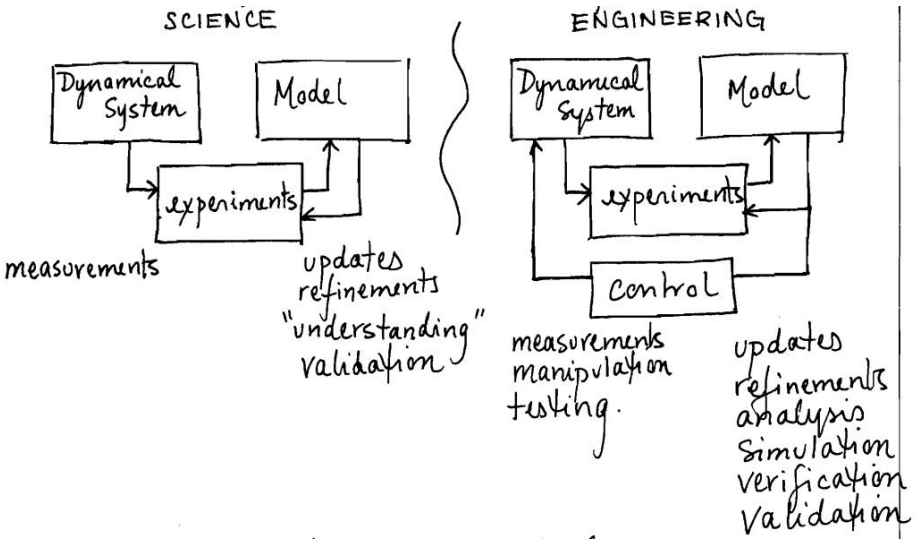
Goals of Lecture 1:

1. An introduction to the broad concepts of modeling and analysis of engineering systems—Modeling, Analysis and control, Verification, Simulation, Validation

2. An overview of the course

The difference between science and engineering is sometimes expressed in terms of interaction with a physical phenomenon. Physical sciences study the phenomenon, while the engineering disciplines design, manipulate, and control the phenomenon. Simply put, scientists describe while engineers control.

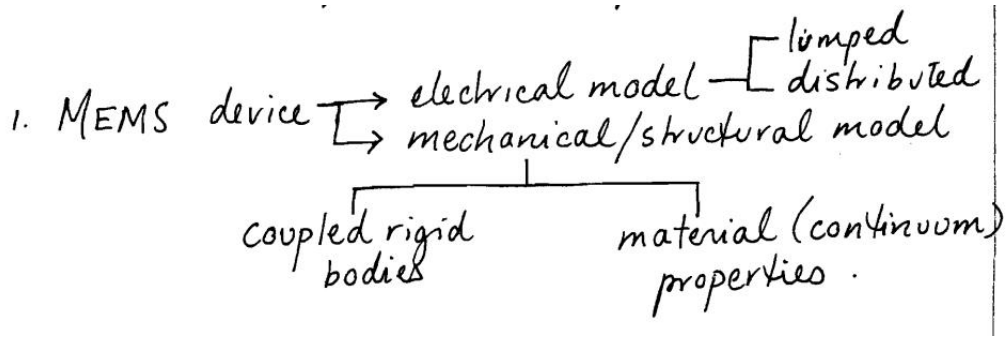
The main purpose of control is to choose an input $u(t)$ to a system such that some pre-defined reward or cost is optimized.



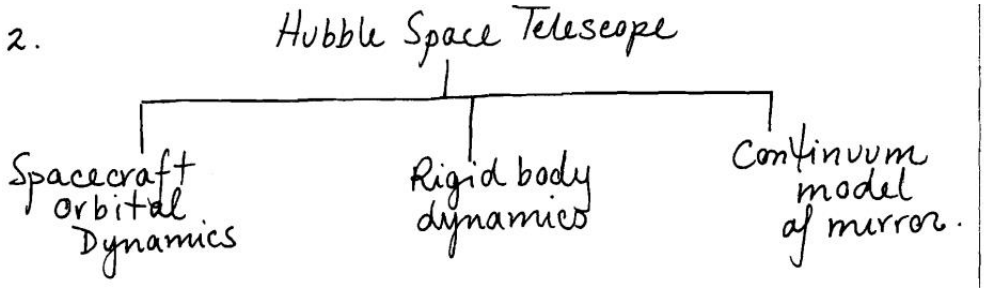
1. **Modeling**:

   The same physical system may have different models, the best choice depends on the problem at hand:

   - MEMS Device:



   - Hubble Space Telescope:



     Which model makes the most sense to use to move the telescope from one altitude to another?

   The utility of a model is in its predictive power: the ability to use it to forecast what the system will do. For example, if, while modeling a system, you notice that input u_i produces output y_i for each $i = 1, \cdots, m$, then a model will be judged by what it predicts the outcome will be when the input is other than $u_1, \cdots, u_n$.

   A basic tenet of science and engineering is that it is of value to find compact, abstract principles which serve as models rather than an exhaustive listing of rules. The principle is that abstraction saves time and effort. Models can then be analyzed, starting from abstract representations.

   In the current age of artificial intelligence (AI) and machine learning (ML), it has become possible to construct more complicated models, and extend (generalize) models to accommodate an infinite number of possible inputs.

   In general, dynamical systems can be categorized by their temporal evolution, linearity, and time invariance.

(a) **Temporal Evolution:**

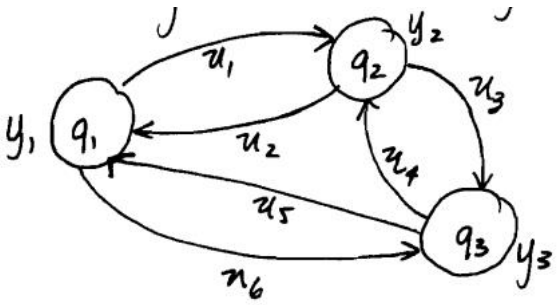
t is a "privileged" variable representing the progression of time. The reason t is "privileged" is for its uni-directionality of evolution. However, there are many models for the evolution of time.

- Continuous time:
$$t \in \mathbb{R}$$
- Discrete time, synchronous:
$$t \in \{nT | n \in \mathbb{Z}\}$$
- Discrete time, asynchronous:
$$t \in \{t_i | i \in \mathbb{N}\}, \text{ or } t \in \mathbb{Z},$$

i.e. e_i represents events in a processor when it receives packets from a bus:

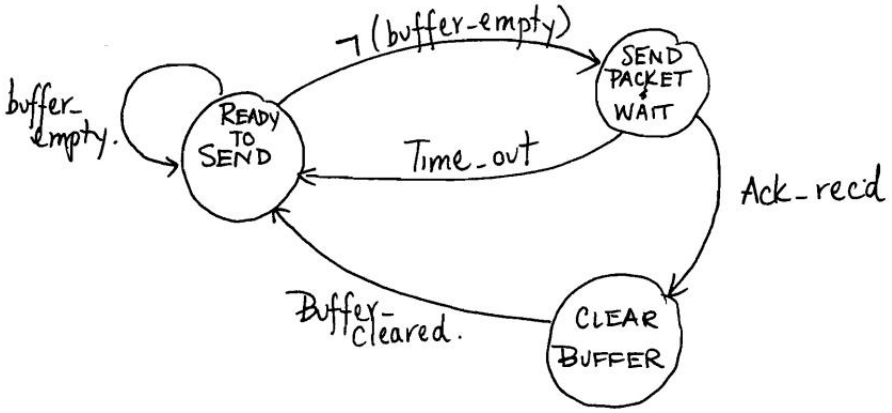
(b) **Linear models, Nonlinear models**:

Suppose the system can be in a finite number of "states." When this "state space" is finite, the system is usually modeled by a *finite state automation*:

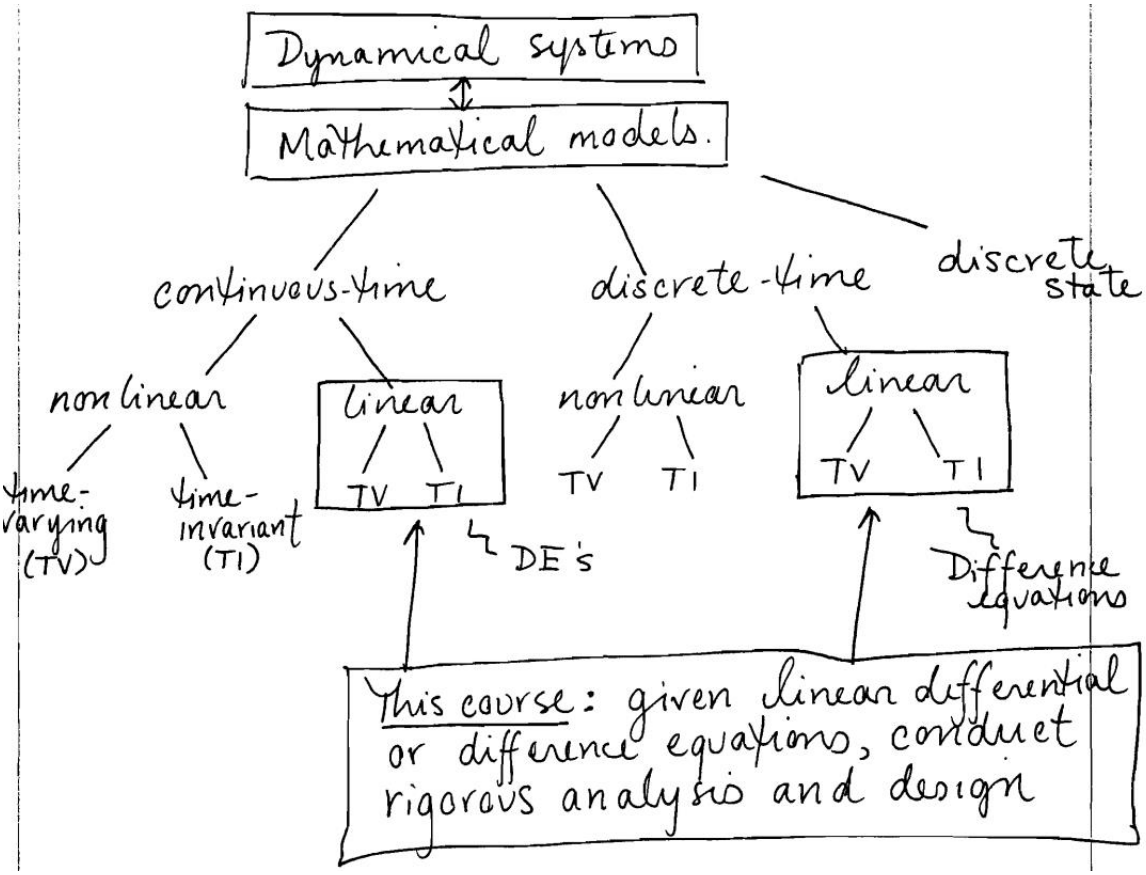


Here, it is understood that u causes the system to transition from state q_1 to state q_2, and so on.

Example. A packet transmitting node transmits a packet and then waits for up to T seconds (timeout) before responding. If an acknowledgement is received, it sends another packet if one is available.



Here, there is no notion of linearity or nonlinearity explicitly (though we can have finite state spaces which have linear structure). When the state space has a *vector space structure*, we can talk about linearity or non-linearity.



Analysis of linear models is much easier than that of nonlinear models (see EECS 222, Spring Semester).

2. **Analysis and Control:**

   Control, broadly speaking, is pervasive; consider the following list of applications:

   - Autopilots on aircraft and spacecraft
   - Air Traffic Control
   - Chemical Process Control
   - Mechatronic:

     Control of ELectromechanical devices, e.g. robots, MEMS, spacecraft, disk drives
   - Power Systems—Generators, turbines, power networks
   - Communication Networks:

     Control of queues, flow control at the network level (congestion management)

- Quantum Chemistry:
  Bond-selective chemistry, e.g. breaking strong bonds and retaining weak bonds
- Biological motor control:
  How we organize control of our musculo-skeletal systems
- Telepresence:
  Projecting your influence to a remote location
- Computational biology:
  Understanding developmental biology

Indeed, some factors that have contributed to a renaissance in control techniques include:

- Cheap processing ability
- CAD tools for inexpensive prototyping new controller designs
- The need to take systems—designed in an era in which suboptimal operation was adequate—to a level of greater performance, e.g. power systems, highways, air traffic, and even the Internet

3. **Verification**:

   Here, one applies control laws or analysis techniques to a possibly more detailed model of the system than the one it was based on to see if one can *analytically verify* or prove that the control law does what is advertised. For example, if you want to regulate some system variables, you shoudl try to prove stability of the system. If you know that a certain region of the state space is unsafe, you need to prove that you can keep the trajectory out of that region.

4. **Simulation**:

   Sometimes, one does not have a clear sense of how all the different parts of the system behave when one focuses on simplified models and objectives. Simulation helps one to get an overall sense of the system design, as well as a sense of how the controller behaves on a model closer to the true system.

5. **Validation**:

   Validation, usually through extensive testing, refers to the act of trying out a design on teh true system and testing if requirements (safety, performance) are met. Since the true system is different from the model, which has gone through several levels of simplification, it is important that the design be *robust*, i.e. its performance degrades continuously ("gently") with changes in the model. A major difficulty lies in the fact that in simulations, only a finite number of states can be sampled at any given time.

   In particular, in the field of *adaptive control*, controllers are designed to adapt to a control system with parameters that vary.

Chapter 2

Linear Algebra Review

Note. In the following, aside from Professor Claire Tomlin's notes [10], we will also make use of definitions, theorems, and examples from Professor Chee-Fai Yung's text on linear algebra [11].

2.1 Lecture 2

Goals of Lecture 2:
Review of Linear Algebra

1. Functions:

   Injective, surjective, bijective, left inverse, right inverse

2. Field, ring

3. Vector space, subspace

4. Linear independence and dependence

5. Basis, coordinates

Note (**Reference**). Callier and Desoer (C;D), Appendix A (A.1 - A.3)

The purpose of abstraction is to describe mathematical, scientific, or engineering concepts that share properties with familiar mathematical objects ($\mathbb{Z}, \mathbb{R},$, etc.)

The following chart describes how linear algebraic concepts are used in the description and analysis of linear systems:

1. Linear (vector) space:

   State space, input space, output space

2. Linear maps:

   Reachability map L_r, Observability map L_o

3. Normed spaces:

   Stability analysis

4. Inner product, Adjoint:

   "Adjoint" linear systems, controllability Grammians, observability Grammians

Recall the notations used for the complex numbers, real numbers, and their subsets $(\mathbb{N}, \mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C})$, the set of quantifiers $(\in, \notin, \forall, \exists, \exists!, \exists?, \ni)$, and implications $(\Longrightarrow, \Longleftarrow, \Longleftrightarrow)$

Definition 2.1 (**Cartesian Product**). *Given two sets $\mathcal{X}$ and $\mathcal{Y}$, the Cartesian product $\mathcal{X} \times \mathcal{Y}$ is the set of all ordered pairs (x, y) such that $x \in \mathcal{X}$ and $y \in \mathcal{Y}$:*

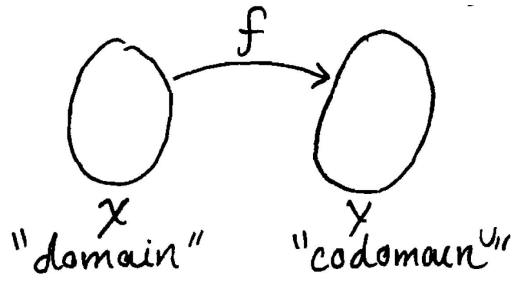
$$\mathcal{X} \times \mathcal{Y} \equiv \{(x, y) | x \in \mathcal{X}, y \in \mathcal{Y}\}$$

The set of all ordered n-tuples of real (complex) numbers is denoted by $\mathbb{R}^n$ ($\mathbb{C}^n$).

Definition 2.2 (**Functions**). *Given two sets $\mathcal{X}$ and $\mathcal{Y}$, by $f : \mathcal{X} \to \mathcal{Y}$, we mean that for each $x \in \mathcal{X}$, we assign a* unique *$f(x) \in \mathcal{Y}$. In this case, we say that f maps the **domain** $\mathcal{X}$ **into** the **codomain** (or **image domain**) $\mathcal{Y}$, and we say that:*

$$f(\mathcal{X}) = \{f(x) | x \in \mathcal{X}\}$$

*is the **range** of f.*



Definition 2.3 (**Injective Function**). *A function $f : \mathcal{X} \to \mathcal{Y}$ is called **injective** (or **one-to-one**, or **1-1**) if $f(x_1) = f(x_2)$ for any $x_1, x_2 \in \mathcal{X}$, then $x_1 = x_2$. In other words:*

$$\begin{aligned} f \text{ is injective} &\Longleftrightarrow f(x_1) = f(x_2) \Rightarrow x_1 = x_2 \\ &\Longleftrightarrow x_1 \neq x_2 \Rightarrow f(x_1) \neq f(x_2) \end{aligned}$$

Definition 2.4 (**Surjective Function**]). *A function $f : \mathcal{X} \to \mathcal{Y}$ is called **surjective** if, for each $y \in \mathcal{Y}$, there exists some $x \in \mathcal{X}$ such that $y = f(x)$; in other words:*

$$\begin{aligned} f \text{ is surjective} &\Longleftrightarrow \forall y \in \mathcal{Y}, \exists x \in \mathcal{X} \ni y = f(x) \\ &\Longleftrightarrow f(\mathcal{X}) = \mathcal{Y} \end{aligned}$$

Definition 2.5 (**Bijective Function, Bijection**). *A function $f : \mathcal{X} \to \mathcal{Y}$ is called **bijective**, or a **bijection**, if it is both injective and surjective; in other words:*

$$f \text{ is bijective} \iff \forall y \in \mathcal{Y}, \exists! x \in \mathcal{X} \ni y = f(x)$$

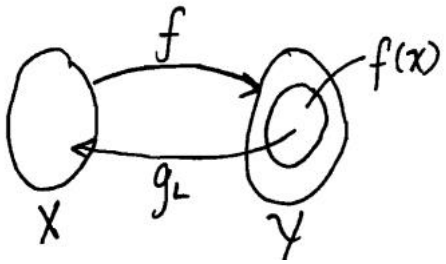
Definition 2.6 (**Left and Right Inverse**). *Consider $f : \mathcal{X} \to \mathcal{Y}$, and let $\mathbf{1}_{\mathcal{X}}$ be the identity map on $\mathcal{X}$. We define the **left inverse** of f,* if it exists, *as the map $g_L : \mathcal{Y} \to \mathcal{X}$ such that*

$$g_L \circ f = \mathbf{1}_{\mathcal{X}},$$

*where $g_L \circ f : X \to \mathcal{X}$, the **composition of g_L and f**, is defined by $(g_L \circ f)(x) \equiv g_L(f(x))$.*

*Similarly, we define the **right inverse** of f as the map $g_R : Y \to \mathcal{X}$ such that $f \circ g_R = 1_{\mathcal{Y}}$.*

*If $\mathcal{X} = \mathcal{Y}$, and the left and right inverses of $f : \mathcal{X} \to \mathcal{X}$ exist, then they are identical (Exercise). This mapping is called the **inverse** of f and is denoted as f^{-1}.*



Theorem 2.7. *A function $f : \mathcal{X} \to \mathcal{Y}$ has a left inverse $g_L \iff f$ is injective.*

Proof.

" $\Leftarrow$ " If f is injective, then for each distinct $x_1, x_2 \in \mathcal{X}$, we have $f(x_1) \neq f(x_2)$. Construct a function $g_L : f(\mathcal{X}) \to \mathcal{X}$ such that $g_L(f(x)) = x$ for each $x \in \mathcal{X}$. (This is a well-defined function, due to the injectivity of f). By definition:

$$\begin{aligned} &\because (g_L \circ f)(x) = g_L(f(x)) = x, \quad \forall x \in \mathcal{X} \\ &\therefore g_L \circ f = 1_{\mathcal{X}} \end{aligned}$$

If f has a left inverse g'_L, then by definition of left inverse, $g'_L(f(x)) = x$ for each $x \in \mathcal{X}$, so $g'_L = g_L$. This demonstrates the uniqueness of the left inverse.

" $\Rightarrow$ " Suppose $f : \mathcal{X} \to \mathcal{Y}$ has a left inverse g_L, and let $x_1, x_2 \in \mathcal{X}$ be given such that $f(x_1) = f(x_2)$. Then:

$$x_1 = g_L(f(x_1)) = g_L(f(x_2)) = x_2$$

■

Exercise. Prove that $f : \mathcal{X} \to \mathcal{Y}$ has a right inverse $g_R \iff f$ is surjective.

Definition 2.8 (**Field**). *A **field** is a composed of a set $\mathbb{F}$ of scalars (numbers), on which the two operations addition "$+$" and multiplication "$\cdot$" are defined. These two operations satisfy the properties below:*

1. **Closure of** $+, \cdot$ —*For each $a, b \in \mathbb{F}$, $a + b \in \mathbb{F}$, and $a \cdot b$ (often simplified as ab) $\in \mathbb{F}$*
2. **Commutativity over** $+, \cdot$—*For each $a, b \in \mathbb{F}$, $a + b = b + a$ and $ab = ba$.*
3. **Associativity**—*For each $a, b, c \in \mathbb{F}$, $(a + b) + c = a + (b + c)$ and $(ab)c = a(bc)$.*
4. **Distributivity**—*For each $a, b, c \in \mathbb{F}$, $a(b + c) = ab + ac$.*
5. **Unit Elements**—*There exist scalars $0, 1 \in \mathbb{F}$, where $0 \neq 1$, such that for each $a \in \mathbb{F}$, $a + 0 = 0 + a = a$ and $1 \cdot a = a \cdot 1 = a$.* (It is possible to prove that 0 and 1 are unique).
6. **Additive Inverse**—*For each $a \in \mathbb{F}$, there exists a scalar $b \in \mathbb{F}$ such that $a+b = b+a = 0$.* (It is not difficult to show that b is unique. b is said to be the **additive inverse** of a, and is often denoted as $-a$).
7. **Multiplicative Inverse**—*For each $a \in \mathbb{F}$, where $a \neq 0$, there exists some $c \in \mathbb{F}$ such that $ac = ca = 1$.* (It is possible to prove that c is unique. c is said to be the **multiplicative inverse** of a, and is often denoted as $1/a$ or a^{-1}).

Remark. A motivation for not allowing the additive identity 0 to have a multiplicative inverse is to render the additive and multiplicative identities distinct. In other words, if 0 has a multiplicative inverse, then $0 = 1$.

Example. Recall that $\mathbb{R}, \mathbb{C}$ are the set of real and complex numbers, respectively. Define:

$$\begin{aligned}
\mathbb{R}(s) &\equiv \text{the set of rational functions in } s\text{with coefficients in } \mathbb{R} \\
\mathbb{C}(s) &\equiv \text{the set of rational functions in } s\text{with coefficients in } \mathbb{C} \\
\mathbb{R}[s] &\equiv \text{the set of polynomials in } s\text{with coefficients in } \mathbb{R} \\
\mathbb{R}_{p,o}(s) &\equiv \text{the set of strictly proper rational functions}
\end{aligned}$$

Note that $\mathbb{R}, \mathbb{C}, \mathbb{R}(s), \mathbb{C}(s)$ are fields, while $\mathbb{R}[s], \mathbb{R}_{p,o}(s)$ are not (Not all elements in the latter two sets have multiplicative inverses).

Furthermore, the *subtraction* of two scalars $a, b \in \mathbb{F}$ (i.e. $a - b$) is defined as the addition of a to the additive inverse of b, i.e. $-b$. Similarly, the *division* of two scalars $a, b \in \mathbb{F}$ (i.e. $\frac{a}{b}$), where $b \neq 0$, is defined as the multiplication of a to the multiplicative inverse of b, i.e. b^{-1}.

Below, we introduce several categories of algebraic structures which share properties with the field. The last of these, the vector space, is the most important.

Definition 2.9 (**Ring, Commutative Ring**). *A **ring** is a set that shares all characteristics of a field, except:*

1. *It may not be commutative under $\cdot$*

2. *It may not have an inverse under $\cdot$ for non-zero elements.*

Rings that are commutative under $\cdot$ are known as ***commutative rings****; rings that are not are called* ***non-commutative rings****.*

Example. Examples of commutative rings include:

$$\mathbb{Z}, \mathbb{R}[s], \mathbb{C}[s], \mathbb{R}_{p,o}(s), \mathbb{R}_p(s)$$

Examples of non-commutative rings include:

$$\mathbb{R}^{n\times n}, \mathbb{C}^{n\times n}, \mathbb{R}^{n\times n}[s], \mathbb{C}^{n\times n}[s], \mathbb{R}^{n\times n}(s), \mathbb{C}^{n\times n}(s), \mathbb{R}_p^{n\times n}(s)$$

Definition 2.10 (**Vector Space**)**.** *Let there be a set $\mathcal{V}$ and a field $\mathbb{F}$. If there exist two operations,* ***vector addition*** *$+ : \mathcal{V} \times \mathcal{V} \to \mathcal{V}$ and* ***scalar multiplication*** *$\cdot : \mathbb{F} \times \mathcal{V} \to \mathcal{V}$ (For $a \in \mathbb{F}$, $\boldsymbol{x} \in \mathcal{V}$, $a \cdot \mathbf{x}$ is often written as $a\mathbf{x}$) that satisfy the following axioms, $\mathcal{V}$ is said to be a* ***vector space over the field*** *$\mathbb{F}$:*

1. *For each $\mathbf{x}, \mathbf{y} \in \mathcal{V}$, $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$.*
2. *For each $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{V}$, $(\mathbf{x} + \mathbf{y}) + \mathbf{z} = \mathbf{x} + (\mathbf{y} + \mathbf{z})$.*
3. *There exists an element $\mathbf{0} \in \mathcal{V}$ such that for each $\mathbf{x} \in \mathcal{V}$, $\mathbf{x} + \mathbf{0} = \mathbf{0} + \mathbf{x} = \mathbf{x}$.*
4. *For each $x \in \mathcal{V}$, there exists a $y \in \mathcal{V}$ such that $\mathbf{x} + \mathbf{y} = \boldsymbol{y} + \boldsymbol{x} = \boldsymbol{0}$* ($\mathbf{y}$ is called the *additive inverse* of $\mathbf{x}$).
5. *For each $\mathbf{x} \in \mathcal{V}$ and $a, b \in \mathbb{F}$, $a(b\mathbf{x}) = (ab)\mathbf{x}$.*
6. *For each $\mathbf{x}$, $\mathbf{y} \in \mathcal{V}$ and $a \in \mathbb{F}$, $a(\mathbf{x} + \mathbf{y}) = a\mathbf{x} + b\mathbf{y}$*
7. *For each $\mathbf{x} \in \mathcal{V}$ and $a, b \in \mathbb{F}$, $(a + b)\mathbf{x} = a\mathbf{x} + b\mathbf{x}$*
8. *For each $\mathbf{x} \in \mathcal{V}$, $1 \cdot \mathbf{x} = \mathbf{x}$.*

A vector space $\mathcal{V}$ defined over the field $\mathbb{F}$ is often denoted as $(\mathcal{V}, \mathbb{F})$, or simply ${}_{\mathbb{F}}\mathcal{V}$; it is also common to remove the "$\mathbb{F}$" and simply say that $\mathcal{V}$ is a vector space. The elements of $\mathcal{V}$ are known as *vectors*, while the elements of $\mathbb{F}$ are known as *scalars*. When $\mathbb{F} = \mathbb{R}$ (or $\mathbb{C}$), $\mathcal{V}$ is said to be a *real vector space* (or respectively, *complex vector space*).

The definition of the operator $+ : \mathcal{V} \times \mathcal{V} \to \mathcal{V}$ indicates that $\mathcal{V}$ must be closed under vector addition, that is, for each $\mathbf{x}$, $\mathbf{y} \in \mathcal{V}$, $\mathbf{x} + \mathbf{y} \in \mathcal{V}$. Similarly, the definition of the operator $\cdot : \mathbb{F} \times \mathcal{V} \to \mathcal{V}$ indicates that $\mathcal{V}$ must be closed under scalar multiplication, that is, for each $\mathbf{x} \in \mathcal{V}$ and $a \in \mathbb{F}$, $a\mathbf{x} \in \mathcal{V}$. The first four axioms in Definition 2.5 are related to vector addition, while the latter four are related to scalar multiplication. The first and second axioms respectively describe the *commutative and associative properties of vector addition*. Similarly, the fifth axiom describes the *associative property of scalar multiplication*. The third axiom explains that each vector space consists of at least one vector—the zero vector. The fourth axiom declares that an additive inverse exists for each vector in a vector space. The sixth and seventh axioms describe the *distributive properties of scalar multiplication*. The eighth and last axiom (*identity*), may appear trivial at first glance, but will later be revealed to have great significance.

Example. Examples of vector spaces include:

1. $(\mathbb{F}^n, \mathbb{F})$: The space of n-tuples in $\mathbb{F}$ over the field $\mathbb{F}$. Common examples include $(\mathbb{R}^n, \mathbb{R})$ and $(\mathbb{C}^n, \mathbb{C})$.

2. Consider the function space $F(D, V)$ of all functions which map D to V, where (V, F) is a vector space, D is a set (e.g. $\mathbb{R}, \mathbb{R}^n$, etc.), and each $f, g \in F(D, V), d \in D$ satisfy:

   (a) Addition— $(f + g)(d) = f(d) + g(d)$

   (b) Scalar multiplication— $(\alpha f)d = \alpha\, f(d)$.

   Then $(\mathbb{F}(D, V), \mathbb{F})$ is a vector space.

Examples include:

1. $(\mathscr{C}([t_0, t_1], \mathbb{R}^n), \mathbb{R})$— The set of all continuous functions on $[t_0, t_1] \to \mathbb{R}^n$

2. $(\mathscr{C}^k([t_0, t_1], \mathbb{R}^n), \mathbb{R})$— The set of all k-times differentiable functions on $[t_0, t_1] \to \mathbb{R}^n$

Definition 2.11 (Subspace). *Let $(\mathcal{V}, \mathbb{F})$ be a vector space, and let $\mathcal{W}$ be a subset of $\mathcal{V}$. Apply the vector addition and scalar multiplication operations of $\mathcal{V}$ are applied to $\mathcal{W}$. If $\mathcal{W}$ then becomes a vector space itself, then $\mathcal{W}$ is said to be a **subspace** of $\mathcal{V}$.*

How can one determine whether or not a subset of $\mathcal{V}$ is a subspace? The theorem below describes a simple method for determining the basic properties of a subspace.

Theorem 2.12. *Let $(\mathcal{V}, \mathbb{F})$ be a vector space, and let $\mathcal{W}$ be a subset of $\mathcal{V}$. Then $(\mathcal{W}, \mathbb{F})$ is a vector subspace of $(\mathcal{V}, \mathbb{F})$ if and only if:*

1. $\mathbf{0} \in \mathcal{W}$
2. *For each* $\mathbf{x}, \mathbf{y} \in \mathcal{W}$, $\mathbf{x} + \mathbf{y} \in \mathcal{W}$
3. *For each* $\mathbf{x} \in \mathcal{W}$ *and* $a \in \mathbb{F}$, $a\mathbf{x} \in \mathcal{W}$

where $\mathbf{0}$ is the zero vector of $\mathcal{V}$.

Proof.

Necessity: Suppose $\mathcal{W}$ is a subspace of $\mathcal{V}$. From the closure of vector addition and scalar multiplication, conditions 2 and 3 must be true. Since $\mathcal{W}$ is a vector space, the third axiom of vector spaces implies that there exists in $\mathcal{W}$ a zero vector $0' \in \mathcal{W} \subset \mathcal{V}$, such that for each $\mathbf{x} \in \mathcal{W}$, $\mathbf{x} + \mathbf{0'} = \mathbf{0'} + \mathbf{x} = \mathbf{x}$. But for each $x \in \mathcal{V}$, $\mathbf{x} + \mathbf{0} = \mathbf{0} + \mathbf{x} = \mathbf{x}$, so $\mathbf{x} + \mathbf{0'} = \mathbf{x} + \mathbf{0}$. From the Cancellation Law of Vector Addition, $0' = 0$. In other words, *the zero vector* $\mathbf{0'} \in \mathcal{W}$ *and the zero vector* $\mathbf{0} \in \mathcal{V}$ *are the same.*

Sufficiency: Suppose $\mathcal{W}$ is some subset of $\mathcal{V}$ satisfying the above three conditions. A careful observation of the axioms of the vector field reveals that besides Axiom 4, all axioms that are valid in $\mathcal{V}$ will naturally also be true in $\mathcal{W}$. In fact, Axiom 4 is also valid in W: For each $\mathbf{x} \in \mathcal{W}$, its additive inverse in $\mathcal{V}$ is $-\mathbf{x} = (-1)\cdot\, \mathbf{x} \in \mathcal{W}$. ■

Definition 2.13 (**Linear Dependence and Linear Independence**). *Let $(\mathcal{V}, \mathbb{F})$ be a vector space, and let $\mathcal{S}$ be a subset of $\mathcal{V}$. Then:*

1. *If there exist a finite number of vectors $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$, and corresponding scalars $a_1, a_2, \cdots a_n \in \mathbb{F}$, not all zero, such that:*
$$a_1\mathbf{x}_1 + \cdots + a_n\mathbf{x}_n = 0$$
*then $\mathcal{S}$ is said to be a **linearly dependent subset** of $\mathcal{S}$ (or simply, $\mathcal{S}$ is **linearly dependent**). Likewise, the elements of $\mathcal{S}$ are said to be linearly dependent.*

2. *If $\mathcal{S}$ is not linearly dependent, it is said to be a **linearly independent subset** of $\mathcal{V}$ (or simply, $\mathcal{S}$ is **linearly independent**). Likewise, the elements of $\mathcal{S}$ are said to be linearly independent.*

From the above definition, it is clear that $\mathcal{S}$ is linearly independent if and only if, for each finite subset of arbitrary vectors $\mathbf{x}_1, \cdots, \mathbf{x}_n$ in $\mathcal{S}$:

$$a_1\mathbf{x}_1 + \cdots + a_n\mathbf{x}_n = 0$$

implies $a_1 = a_2 = \cdots = a_n = 0$. Equivalently, $\mathcal{S}$ is linearly independent if and only if the only linear combination of elements of $\mathcal{S}$ that generates the zero vector $\mathbf{0}$ is:

$$\mathbf{0} = 0\mathbf{x}_1 + \cdots + 0\mathbf{x}_n$$

Definition 2.14 (**Basis**). *Let $\mathcal{V}$ be a vector space, and let $\mathcal{B}$ be a subset of $\mathcal{V}$. If:*

1. *$\mathcal{B}$ is a linearly independent subset,*

2. *$\mathcal{B}$ is a spanning subset,*

*then $\mathcal{B}$ is said to be a **basis** for $\mathcal{V}$.*

Theorem 2.15. *Let $\mathcal{V}$ be a vector space, and let $\mathcal{B}$ be a subset of $\mathcal{V}$. Then $\mathcal{B}$ is a basis for $\mathcal{V}$ if and only if each element of $\mathcal{V}$ can be expressed as a unique linear combination of elements of $\mathcal{B}$.*

The unique linear combination of a vector $\mathbf{x}$ with respect to $\mathcal{B}$ is called the **set of coordinates of $\mathbf{x}$ with respect to $\mathcal{B}$**. The coordinate transformation of $\mathbf{x}$ from the basis $\mathcal{B}$ to a different basis $\mathcal{B}'$, is left as an exercise.

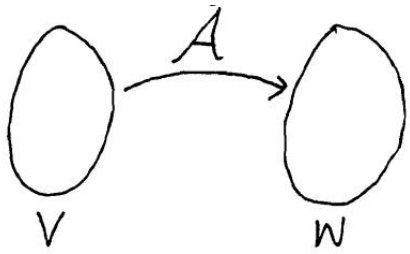
2.2 Lecture 3

This lecture concerns a review of linear transformations and linear operators.

Definition 2.16 (**Linear Transformation**). *Let* $(\mathcal{U}, \mathbb{F})$ *and* $(\mathcal{V}, \mathbb{F})$ *be two vector spaces. If a mapping* $\boldsymbol{L}$*:* $\mathcal{U} \to \mathcal{V}$ *satisfies the following two conditions:*

1. *For each* $\mathbf{x}, \mathbf{y} \in \mathcal{U}$, $\mathbf{L(x + y) = Lx + Ly}$.
2. *For each* $\mathbf{x} \in \mathcal{U}$ *and each* $a \in \mathbb{F}$, $\mathbf{L}(a\mathbf{x}) = a\mathbf{Lx}$.

then L *is said to be a* ***linear map*** *or* ***linear transformation****. If* $\mathcal{U} = \mathcal{V}$*, the linear transformation is also called a* ***linear operator****.*



In the following text, the field $\mathbb{F}$ is often omitted when it is clear that $\mathcal{I}$ and $\mathcal{V}$ are constructed from the same field.

This property can clearly be extended to any number of coefficients, i.e. for any linear transformation $L : \mathcal{U} \to \mathcal{V}$ given $\alpha_1, \cdots, \alpha_n \in \mathbb{F}$ and $v_1, \cdots, v_n \in (\mathcal{V}, \mathbb{F})$, we have:

$$L\left(\sum_{i=1} \alpha_i v_i\right) = \sum_{i=1}^{n} \alpha_i L(v_i)$$

Example. Consider the mapping:

$$L : as^2 + bs + c \longrightarrow \int_0^s (bt + a)dt$$

Then, with respect to the basis $\mathcal{B} = \{s^2, s, 1\}$, the representation of the mapping can be described as follows:

$$(a, b, c) \longrightarrow \left(\frac{1}{2}b, a, 0\right)$$

Clearly, L is linear.

Example. If we modify the linear mapping shown above to:

$$L' : as^2 + bs + c \longrightarrow \int_0^s (bt + a)dt + 5$$

Then this is no longer a linear map, since the zero vector 0 is no longer mapped to itself. However, since $L' - 5$ is a linear map, L' can be considered the composition of a linear map and a translation. Such maps are known as **affine maps**.

Definition 2.17 (**Range, Null Space**). *Let a linear map $L : \mathcal{U} \to \mathcal{V}$ be given.*

1. *The **range**, or image, of L is the subspace defined by:*

$$R(L) = \{v | v = L(u), u \in \mathcal{U}\} \leq \mathcal{V}$$

2. *The **null space**, or kernel, of L is the subspace defined by:*

$$N(L) = \{u | L(u) = 0\} \leq \mathcal{U}$$

It is not immediately obvious from their definitions that the range and null spaces of a linear map are subspaces; this must be explicitly proved.

The range and null space are critical in the analysis of the injectivity and surjectivity of a linear map. Given a system of linear equations characterized by:

$$Lu = b,$$

where $b \in \mathcal{V}$ is given, the solution exists if L is surjective and is unique if L is both injective and surjective (i.e. if L is bijective).

Theorem 2.18 (**Dimension Theorem**). *Let $\mathcal{U}, \mathcal{V}$ be vector spaces, where where $dim\mathcal{U} = n < \infty$, and let $\mathbf{L} \in \mathcal{L}(\mathcal{U}, \mathcal{V})$. Then:*

$$\text{nullity}(\mathbf{L}) + \text{rank}(\mathbf{L}) = \dim\mathcal{U}. \tag{2.1}$$

Proof. Exercise. ■

Below, we discuss the matrix representations of linear mappings. Let $\mathcal{U}, \mathcal{V}$ be finite-dimensional vector spaces, and let a linear mapping $L : \mathcal{U} \to \mathcal{V}$ be given. Let $\mathcal{B}_\mathcal{U} = \{u_j\}_{j=1}^n$ and $\mathcal{B}_\mathcal{V} = \{v_i\}_{i=1}^m$ be bases for $\mathcal{U}$ and $\mathcal{V}$, respectively. In affect, the task of finding the image, under L, of an infinite number of elements in $\mathcal{U}$ can be reduced to find the image of each element in the finite set $\{u_j\}_{j=1}^n$. First, we associate $\mathcal{U}$ with $\mathcal{V}$ as follows:

$$L(u_j) = \sum_{i=1}^m a_{ij} v_i$$

Then, given any linear combination of $\{u_j\}_{j=1}^n$, e.g. $x = \sum_{j=1}^n \xi_j u_j$, we have:

$$\begin{aligned} L(x) &= \sum_{j=1}^n \xi_j \, L(u_j) = \sum_{j=1}^n \xi_j \sum_{i=1}^m a_{ij} v_i \\ &= \sum_{i=1}^m \left(\sum_{j=1}^n a_{ij} \xi_j \right) v_i = \sum_{i=1}^m \eta_i v_i \end{aligned}$$

where we have defined $\eta_i \equiv \sum_{j=1}^n a_{ij} \xi_j$. In other words:

$$\eta = A\xi$$

where $\eta = (\eta_1, \cdots, \eta_m)$, $\xi = (\xi_1, \cdots, \xi_n)$, and $A = [a_{ij}]_{m\times n}$, i.e.:

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

We conclude that *the linear map is uniquely defined by the matrix A*. We say that A is the **matrix representation** of L with respect to $\mathcal{U}$ and $\mathcal{V}$; this is denoted by:

$$A = [L]_{\mathcal{U}}^{\mathcal{V}}$$

The following is useful to remember—The j^{th} column of the matrix A contains the coordinates of the vector $A(u_j) \in \mathcal{V}$ expressed with respect to $\{v_i\}_{i=1}^m$.

Suppose, given some linear operator $L : \mathbb{R}^n \to \mathbb{R}^n$, there exists a $b \in \mathbb{R}^n$ such that:

$$\mathcal{B}_{\mathcal{V}} = \{b, Lb, \cdots, L^{n-1}b\}$$

forms a basis for $\mathcal{V}$. Then, since $L^n b \in \mathcal{V}$, one can find coordinates $\alpha_1, \cdots, \alpha_n$ such that:

$$L^n b = -\alpha_n b - \alpha_{n-1} Lb - \cdots - \alpha_1 L^{n-1} b$$

In this case, the coordinate representation of b and the matrix representation A of L, both with respect to $\mathcal{B}_{\mathcal{V}}$, are:

$$b = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad A = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & -\alpha_n \\ 1 & 0 & 0 & \cdots & 0 & -\alpha_{n-1} \\ 0 & 1 & 0 & \cdots & 0 & -\alpha_{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & -\alpha_2 \\ 0 & 0 & 0 & \cdots & 1 & -\alpha_1 \end{bmatrix}$$

This is often helpful in the analysis of linear systems (e.g. characterized by the form $\dot{x} = Ax + Bu$), in which the entire state space can be decomposed into subspaces with matrix representation of the above form.

Below, we consider the relationship between two matrix representations of the same linear map $L : \mathcal{U} \to \mathcal{V}$. Let $\mathcal{B}_{\mathcal{U}} = \{u_j\}_{j=1}^n$ and $\mathcal{B}_{\overline{\mathcal{U}}} = \{\overline{u_j}\}_{j=1}^n$ be bases for $\mathcal{U}$, and let $\mathcal{B}_{\mathcal{V}} = \{v_i\}_{i=1}^m$ and $\mathcal{B}_{\overline{\mathcal{V}}} = \{\overline{v_i}\}_{i=1}^m$ be bases for $\mathcal{V}$. Let $A, \overline{A}$ be the matrix representations of L with respect to $\mathcal{B}_{\mathcal{U}}, \mathcal{B}_{\mathcal{V}}$ and $\mathcal{B}_{\overline{\mathcal{U}}}, \mathcal{B}_{\overline{\mathcal{V}}}$, respectively.

Now, let $x \in \mathcal{U}$ be given, and suppose:

$$x = \sum_{i=1}^{m} \xi_i u_i = \sum_{i=1}^{m} \overline{\xi_i}\,\overline{u_i}$$

$$\Rightarrow x = \begin{bmatrix} u_1 & u_2 & \cdots & u_n \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_n \end{bmatrix} = \begin{bmatrix} \overline{u_1} & \overline{u_2} & \cdots & \overline{u_n} \end{bmatrix} \begin{bmatrix} \overline{\xi_1} \\ \overline{\xi_2} \\ \vdots \\ \overline{\xi_n} \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_n \end{bmatrix} = \underbrace{\begin{bmatrix} u_1 & u_2 & \cdots & u_n \end{bmatrix}^{-1} \begin{bmatrix} \overline{u_1} & \overline{u_2} & \cdots & \overline{u_n} \end{bmatrix}}_{\equiv P} \begin{bmatrix} \overline{\xi_1} \\ \overline{\xi_2} \\ \vdots \\ \overline{\xi_n} \end{bmatrix} = P \begin{bmatrix} \overline{\xi_1} \\ \overline{\xi_2} \\ \vdots \\ \overline{\xi_n} \end{bmatrix}$$

Similarly, for $Lx \in \mathcal{V}$, we have:

$$Lx = \sum_{i=1}^{m} \eta_i v_i = \sum_{i=1}^{m} \overline{\eta_i v_i}$$

$$\Rightarrow x = \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_n \end{bmatrix} = \begin{bmatrix} \overline{v_1} & \overline{v_2} & \cdots & \overline{v_n} \end{bmatrix} \begin{bmatrix} \overline{\eta_1} \\ \overline{\eta_2} \\ \vdots \\ \overline{\eta_n} \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_n \end{bmatrix} = \underbrace{\begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix}^{-1} \begin{bmatrix} \overline{v_1} & \overline{v_2} & \cdots & \overline{v_n} \end{bmatrix}}_{\equiv Q} \begin{bmatrix} \overline{\eta_1} \\ \overline{\eta_2} \\ \vdots \\ \overline{\eta_n} \end{bmatrix} = Q \begin{bmatrix} \overline{\eta_1} \\ \overline{\eta_2} \\ \vdots \\ \overline{\eta_n} \end{bmatrix}$$

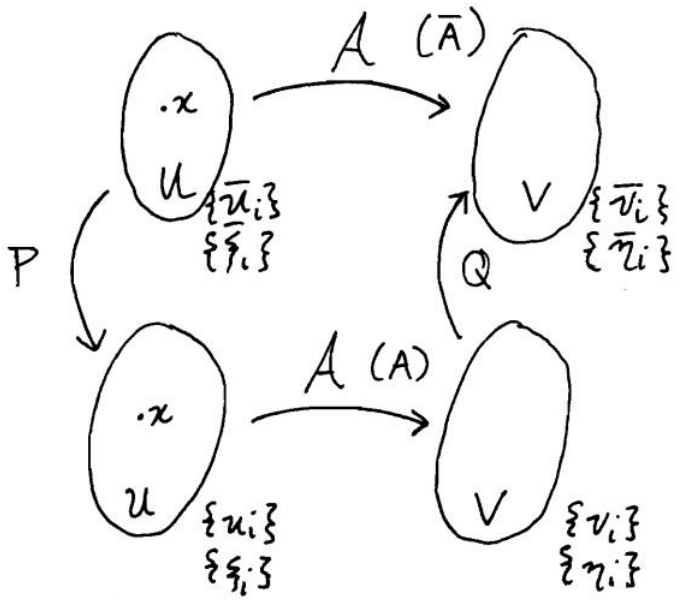
Now, if we define $A \equiv [L]_{\mathcal{U}}^{\mathcal{V}}$, then:

$$\begin{aligned} &\because \eta = A\xi \\ &\therefore \overline{\eta} = Q^{-1}\eta = Q^{-1}A\xi = Q^{-1}AP\overline{\xi}, \end{aligned}$$

where:

$$[L]_{\mathcal{B}_{\overline{\mathcal{U}}}}^{\mathcal{B}_{\overline{\mathcal{V}}}} \equiv \overline{A} = Q^{-1}AP$$

is the matrix representation of A with respect to $\mathcal{B}_{\overline{\mathcal{U}}}\{\overline{u_i}\}, \mathcal{B}_{\overline{\mathcal{V}}}\{\overline{v_j}\}$.



Essentially, a change in basis is equivalent to a change in matrix or vector representation.

Definition 2.19 (**Rank, Nullity**). *Given a linear mapping $L : \mathcal{U} \longrightarrow \mathcal{V}$, define:*

1. $\text{rank}(L) = \dim(R(L))$.
2. $\text{nullity}(L) = \dim(N(L))$.

We conclude our discussion on linear transformations with a corollary that summarizes the important relationships between injectivity, surjectivity, and the existence of right and left inverses.

Theorem 2.20. *Let V, W be vector spaces, with dimensions m, n and bases $\mathcal{B}_V, mathcalB_W$, respectively. Let $L : V \to W$ be a mapping from V to W, with associated matrix representation $A \equiv L_{\mathcal{B}_V}^{\mathcal{B}_W} \in \mathbb{R}^{m \times n}$. Then:*

1. *The following statements are equivalent:*
   - *L is injective.*
   - *L is left invertible.*
   - *$N(L) = 0_V$.*
   - *A has full column rank.*
2. *The following statements are equivalent:*
   - *L is surjective.*
   - *L is right invertible.*

- $R(L) = W$.
- *A has full row rank.*

Proof.

1. "(1) $\Rightarrow$ (2)" : Suppose L is injective. Then, by definition, for each $w \in R(L)$, one can identify a unique $v \in V$ such that $Lv = w$. The left inverse of L can thus be defined as the mapping that associates each $w \in R(L)$ with the unique $v \in \mathcal{V}$ such that $Lv = w$.

   "(2) $\Rightarrow$ (3)" : Suppose there exists some left inverse L^{-1} of L, and let $v \in N(L)$ be arbitrarily given, i.e. $Lv = 0$. Then, applying L^{-1} to the *left* on both sides, we have:

$$v = L^{-1}0 = 0$$

   "(3) $\Rightarrow$ (4)" : Suppose $N(L) = 0_V$. Then $N(A) = 0_{\mathbb{R}^n}$. This implies that, if A_i denotes the i-th column of A, for each $i = 1, \cdots, n$, and a vector $a \equiv (a_1, \cdots, a_n)^T \in \mathbb{R}^n$ were given such that:

$$0 = Av = a_1A_1 + \cdots + a_nA_n,$$

   then $a_1 = \cdots = a_n = 0$. This implies that $\{A_1, \cdots, A_n\}$ are linearly independent, i.e A has full column rank.

   "(4) $\Rightarrow$ (1)" : We will demonstrate that (4) $\Rightarrow$ (3) $\Rightarrow$ (1). By reversing the above argument, we see that if A has full column rank, i.e. if its columns are linearly independent, then $0_{\mathbb{R}^n}$ is the only vector contained in $N(A)$, i.e. $N(A) = \{0_{\mathbb{R}^n}\}$, and (3) immediately follows.

   Now, suppose there exist some $v_1, v_2 \in V$ and $w \in W$ such that:

$$Lv_1 = Lv_2 = w.$$

   Then $L(v_1 - v_2) = 0$, and since $N(L) = 0_V$, we have $v_1 - v_2 = 0_V$, or $v_1 = v_2$. This shows that for each $w \in W$, there exists a most one $v \in \mathcal{V}$ such that $L(v) = w$ (specifically, if $w \notin R(L)$, no such v exists; if $w \in R(L)$, then one unique v exists). This establishes (1).

   We have thus shown that (1) $\Rightarrow$ (2) $\Rightarrow$ (3) $\Rightarrow$ (4) $\Rightarrow$ (1), i.e. (1), (2), (3), (4) are equivalent.

2. Simply take the transpose of the above arguments; the desired results follow from the simple fact that, if a matrix has full row rank, its transpose matrix must have full column rank.

■

Corollary 2.21.

2.3 Lecture 3 Discussion

Example (**Discussion 6, Problem 5**). Let $\mathcal{B} = \{e_1, e_2, e_3\}$ be the standard ordered basis in $\mathbb{R}^3$, and let $\mathcal{C}$ be the ordered basis given by:

$$\mathcal{C} = \{c_1, c_2, c_3\} \equiv \left\{ \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \right\}$$

Let $\mathcal{A} : (\mathbb{R}^3, \mathbb{R}) \to (\mathbb{R}^3, \mathbb{R})$ be a linear map, with:

$$\mathcal{A}(e_1) = \begin{bmatrix} 2 \\ -1 \\ 0 \end{bmatrix}, \quad \mathcal{A}(e_2) = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad \mathcal{A}(e_3) = \begin{bmatrix} 0 \\ 4 \\ 2 \end{bmatrix}.$$

Find the following:

1. $[\mathcal{A}]_{\mathcal{B}}^{\mathcal{C}}$, i.e. the matrix representation of $\mathcal{A}$ with respect to basis $\mathcal{B}$ for domain and basis $\mathcal{C}$ for codomain.
2. $[\mathcal{A}]_{\mathcal{C}}^{\mathcal{C}}$, i.e. the matrix representation of $\mathcal{A}$ with respect to basis $\mathcal{C}$ for both domain and codomain.

Solution :

1. Let $A \equiv [A]_{\mathcal{B}}^{\mathcal{B}}$, the matrix representation of A given with respect to the standard basis. The problem gives us:

$$A = \begin{bmatrix} \mathcal{A}e_1 & \mathcal{A}e_2 & \mathcal{A}e_3 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ -2 & 0 & 4 \\ 0 & 0 & 2 \end{bmatrix}$$

   By definition of change of basis, for each $i \in \{1, 2, 3\}$ the elements of the vector $[\mathcal{A}]_{\mathcal{B}}^{\mathcal{C}} e_i$ are the coefficients, in the same ordering, of Ae_i when expressed as a linear combination of basis vectors in $\mathcal{C}$. In other words, if $[\mathcal{A}]_{\mathcal{B}}^{\mathcal{C}} e_i = (a_{1i}, a_{2i}, a_{3i})$, then:

$$Ae_i = a_{1i}b_1 + a_{2i}b_2 + a_{3i}b_3$$

   We thus have:

$$\begin{aligned} \begin{bmatrix} c_1 & c_2 & c_3 \end{bmatrix} [\mathcal{A}]_{\mathcal{B}}^{\mathcal{C}} &= A \begin{bmatrix} e_1 & e_2 & e_3 \end{bmatrix} \\ \Rightarrow [\mathcal{A}]_{\mathcal{B}}^{\mathcal{C}} &= \begin{bmatrix} c_1 & c_2 & c_3 \end{bmatrix}^{-1} A \begin{bmatrix} e_1 & e_2 & e_3 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 2 & 0 & 0 \\ -2 & 0 & 4 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \\ &= \frac{1}{2} \begin{bmatrix} 1 & 0 & 2 \\ -3 & 0 & 6 \\ 3 & 0 & -2 \end{bmatrix} \end{aligned}$$

2. Repeating the above process, we have:

$$
\begin{aligned}
[\mathcal{A}]_{\mathcal{C}}^{\mathcal{C}} &= \begin{bmatrix} c_1 & c_2 & c_3 \end{bmatrix}^{-1} A \begin{bmatrix} c_1 & c_2 & c_3 \end{bmatrix} \\
&= \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 2 & 0 & 0 \\ -2 & 0 & 4 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} \\
&= \frac{1}{2} \begin{bmatrix} 1 & 2 & 3 \\ -3 & 6 & 3 \\ 3 & -2 & 1 \end{bmatrix}
\end{aligned}
$$

2.4 Lecture 4

Definition 2.22 (**Normed Linear Spaces**). *Let the field $\mathbb{F}$ be $\mathbb{R}$ or $\mathbb{C}$. A linear space $(\mathcal{V}, \mathbb{F})$ is said to be a* ***normed linear space*** *if there exists a map:*

$$\|\cdot\| : \mathcal{V} \longrightarrow \overline{\mathbb{R}^+}$$

satisfying the following axioms, for each $v_1, v_2 \in \mathcal{V}$:

1. $\|v_1 + v_2\| \leq \|v_1\| + \|v_2\|$
2. $\|\alpha v\| \leq |\alpha| \cdot \|v\|$
3. $\|v\| = 0$ *if and only if* $v = 0$.

Example. Common norms for the Euclidean space $\mathbb{R}^n$ are, for each $x = (x_1, \cdots, x_n)$:

1. Two-norm:
$$\|x\|_2 = \left(\sum_{i=1}^{\infty} |x_i|^2 \right)^{\frac{1}{2}}$$
2. One-norm:
$$\|x\|_1 = \sum_{i=1}^{\infty} |x_i|$$
3. p-norm:
$$\|x\|_p = \left(\sum_{i=1}^{\infty} |x_i|^p \right)^{\frac{1}{p}}$$
4. ∞-norm:
$$\|x\|_\infty = \max_i |x_i| = \lim_{p \to \infty} \|x\|_p$$

The following is *not* a norm, but is still sometimes referred to as the "l_0-norm"

$$\lim_{p \to 0} \|x\|_p = \text{number of non-zero entries in } x$$

Definition 2.23 (**Equivalent Norms**). *Two norms $\|\cdot\|_a$ and $\|\cdot\|_b$ are called* ***equivalent*** *if there exist $\alpha, \beta \in \mathbb{R}^+$ such that, for each $v \in \mathcal{V}$:*

$$m_l \|v\|_a \leq \|v\|_b \leq m_u \|v\|_a$$

In other words, $\|\cdot\|_a$ and $\|\cdot\|_b$ are topologically equivalent, in that qualitative comparisons of norms of different vectors hold in any norm. *It can be shown that any two norms on a finite-dimensional vector space are equivalent.*

Example. Common norms for the function space $C([t_0, t_1], \mathbb{R}^n)$ are, for each $f \in C([t_0, t_1], \mathbb{R}^n)$:

1. Two-norm:

$$\|f\|_2 = \left[\int_{t_0}^{t_1} \|f(t)\|^2\right]^{\frac{1}{2}}$$

2. ∞-norm:

$$\|f\|_\infty = \max\{\|f(t)\|_\infty, t \in [t_0, t_1]\}$$

Definition 2.24 (**Induced Norms**). *Let $L : (\mathcal{U}, \mathbb{F}) \longrightarrow (\mathcal{V}, \mathbb{F})$ be a continuous linear operator, and suppose $\mathcal{U}, \mathcal{V}$ are endowed with the norms $\|\cdot\|_\mathcal{U}$ and $\|\cdot\|_\mathcal{V}$, respectively. Then the induced norms of L is defined by:*

$$\|A\|_i = \sup_{u \neq 0} \frac{\|Au\|_\mathcal{V}}{\|u\|_\mathcal{U}}$$

Theorem 2.25 (**Facts about Induced Norms**). *Let $(\mathcal{U}, \|\cdot\|_\mathcal{U})$, $(\mathcal{V}, \|\cdot\|_\mathcal{V}), (\mathcal{W}, \|\cdot\|_\mathcal{W})$ be normed linear spaces, and define $L, \tilde{L} : \mathcal{V} \to \mathcal{W}$ and $M : \mathcal{U} \to \mathcal{V}$. Then, for each $v \in \mathcal{V}$ and $\alpha \in \mathbb{R}$:*

1. $\|Lv\|_\mathcal{W} \leq \|L\|_i \cdot \|v\|_\mathcal{V}$
2. $\|\alpha L\| = |\alpha| \cdot \|L\|$
3. $\|L + \tilde{L}\|_i \leq \|L\|_i + \|\tilde{L}\|_i$
4. $\|L\|_i = 0$ *if and only if* $L = 0$.
5. $\|LM\|_i \leq \|L\|_i \cdot \|M\|_i$.

This definition leads naturally to the concept of sensitiviity—A measure of how the solution x to $Ax = b$ changes as A and b are perturbed. Let $A : \mathbb{F}^n \to \mathbb{F}^n$, with $b \in \mathbb{F}^n$. If A^{-1} exists, the solution $x = A^{-1}b$ is unique, and is denoted as the *nominal solution*:

$$x_0 \equiv A^{-1}b$$

Now, suppose A and b undergo the following perturbations:

$$A \longrightarrow A + \delta A$$
$$b \longrightarrow b + \delta b$$

and the solution x_0 undergoes a perturbation $x_0 \to x_0 + \delta x$. Then:

$$\begin{aligned}
&(A+\delta)(x_0+\delta x) = b + \delta b \\
\Rightarrow &Ax_0 + A\delta x + \delta A x_0 \approx b + \delta b \\
\Rightarrow &A\delta x + \delta A x_0 = \delta b \\
\Rightarrow &\delta x = A^{-1}[-\delta A \cdot x + \delta b] \\
\Rightarrow &|\delta x| \leq \|A^{-1}\|_i \cdot \left[\|\delta A\|_i \cdot |x| + |\delta b|\right], \\
\Rightarrow &\frac{|\delta x|}{|x_0|} \leq \|A^{-1}\|_i \cdot \left[\|\delta A\|_i + \frac{|\delta b|}{|x_0|}\right], \\
&\qquad \leq \|A^{-1}\|_i \cdot \|A\| \cdot \left[\frac{\|\delta A\|_i}{\|A\|} + \frac{|\delta b|}{\|A\||x_0|}\right], \\
&\qquad \leq \|A^{-1}\|_i \cdot \|A\| \cdot \left[\frac{\|\delta A\|_i}{\|A\|} + \frac{|\delta b|}{|b|}\right],
\end{aligned}$$

since $Ax_0 = b$ implies $\|A\| \cdot |x_0| \geq |b|$.

The quantity:

$$\kappa(A) \equiv \|A^{-1}\|_i \cdot \|A\| \geq 1$$

is defined as the *condition number* of A. If $\kappa(A) \gg 1$, then small changes in δb and δA can induce large changes in δx. In other words, the larger the conditional number of A is, the more difficult the system is to stabilize.

Definition 2.26 (**Inner Product Space**). *Let the field $\mathbb{F}$ be $\mathbb{R}$ or $\mathbb{C}$, and consider the linear space $(H, \mathbb{F})$. The function:*

$$\langle \cdot, \cdot \rangle : H \times H \longrightarrow \mathbb{R}^{\mathbb{C}}$$

*is called an **inner product** if, for each $x, y, z \in H$ and $\alpha \in \mathbb{F}$:*

1. $\langle x, y+z \rangle = \langle x, y \rangle + \langle x, z \rangle$.
2. $\langle x, \alpha y \rangle = \alpha \langle x, y \rangle$
3. $|x|^2 \equiv \langle x, x \rangle > 0$ *iff* $x \neq 0_H$.
4. $\langle x, y \rangle = \overline{\langle y, x \rangle}$.

A complete inner product space is known as a *Hilbert space*. Here, completeness refers to the fact that every Cauchy sequence in the space must converge.

A vector $v \in H$ that satisfies $|v| = \langle v, v \rangle = 1$ is said to be a *unit vector*.

Example.

1. $(\mathbb{F}^n, \mathbb{F}, \langle \cdot, \cdot \rangle)$ is a Hilbert space under the inner product:

$$\langle x, y \rangle \equiv \sum_{i=1}^{n} \overline{x_i} y_i = x^\star y$$

for each $x, y \in \mathbb{F}^n$.

2. $L_2([t_0, t_1], \mathbb{F}^n)$, the space of square integrable $\mathbb{F}^n$-valued functions on $[t_0, t_1]$, is a Hilbert space under the inner product:

$$\langle f, g \rangle \equiv \int_{t_0}^{t_1} f(t)^\star g(t)\, dt$$

for each $f, g \in L_2([t_0, t_1], \mathbb{F}^n)$.

Inner products allow us to define orthogonality of vectors and discuss angles between vectors. Henceforth, we will abbreviate $(H, \mathbb{F}, \langle \cdot, \cdot \rangle)$ as H.

Definition 2.27 (**Orthogonality**). *If H is an inner product space space, then $x, y \in H$ are said to be **orthogonal** if $\langle x, y \rangle = 0$, and $x \in H$ is said to be orthogonal to a subset $S \subset H$ if $\langle x, s \rangle = 0$ for each $s \in S$.*

Definition 2.28 (**Orthogonal Complement**). *If $(H, \mathbb{F}, \langle \cdot, \cdot \rangle)$ is an inner product space, and $M \subset H$, then:*

$$M^\perp \equiv \{y \in H : \langle x, y \rangle = 0, \forall\, x \in M\} \leq H$$

*is a subspace of H, and is called the **orthogonal complement** of M.*

Theorem 2.29. *Given a Hilbert space H and some $M \subset H$:*

1. $M \cap M^\perp = \{0\}$.
2. $(M^\perp)^\perp = span(M)$.

Proof.

1. Let $x \in H$ be given such that $x \in M \cap M^\perp$. Then, since $x \in M$ and $x \in M^\perp$, we have $\langle x, x \rangle = 0$. This implies that $x = 0_H$.
2. Exercise.

■

Given a Hilbert space H and some subset $M \subset H$, then $H = M \oplus M^\perp$, where "$\oplus$" denotes the direct sum. (In fact, $H = M \overset{\perp}{\oplus} M^\perp$, where "$\oplus$" denotes the orthogonal direct sum.)

Definition 2.30 (**Adjoint**). *Let $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$, and let $(U, F, \langle \cdot, \cdot \rangle_u)$ and $(V, F, \langle \cdot, \cdot \rangle_u)$ be innner product spaces. Let $A : U \to V$ be continuous and linear. Then the* ***adjoint*** *of A, denoted as $A^\star$, is the (unique) map $A^\star : V \to U$ such that:*

$$\langle v, Au \rangle_{\mathcal{V}} = \langle A^\star v, u \rangle_{\mathcal{U}}$$

Example (**From an old prelim**). Let $f(\cdot), g(\cdot) \in \mathscr{C}([t_0, t_1], \mathbb{R}^n)$ and define a linear map $A : \mathscr{C}([t_0, t_1], \mathbb{R}^n) \to \mathbb{R}$ by:

$$A(f(\cdot)) = \langle g(\cdot), f(\cdot) \rangle$$

for any $f(\cdot), g(\cdot) \in \mathscr{C}([t_0, t_1], \mathbb{R}^n)$. Find the adjoint map of A.

Solution :

By definition of adjoint, $A^\star$ must satisfy, for any $x \in \mathbb{R}$, $f, g \in \mathscr{C}([t_0, t_1], R^n)$:

$$\langle A^\star x, f(\cdot) \rangle = \langle Af(\cdot), x \rangle_{\mathbb{R}} = x \cdot \langle g(\cdot), f(\cdot) \rangle$$

In other words, the mapping is characterized by:

$$A^\star x = x \cdot g(\cdot)$$

2.5 Lecture 4 Discussion

Example (**Discussion 8, Problem 2**). Let M, N be two subspaces of V. Suppose $\mathcal{B}_M \equiv \{m_1, \cdots, m_p\}$ and $\mathcal{B}_N\{n_1, \cdots, n_q\}$ form bases for M and N, respectively. Show that $V = M \oplus N$ if and only if $\mathcal{B} = \{m_1, \cdots, m_p, n_1, \cdots, n_q\}$ is a basis of V.

Solution:

" $\Rightarrow$ " Suppose $V = M \oplus N$. To verify that $\mathcal{B}$ is a basis of V, we need to show that the vectors in $\mathcal{B}$ (1) are linearly independent, and (2) generate V.

1. Let $v \in V = M \oplus N$ be given arbitrarily. Then there exists some (unique) $m \in M$ and $n \in N$ such that $v = m + n$. Since $\mathcal{B}_M$ and $\mathcal{B}_N$ are bases for M and N, respectively, there exist some (unique) scalars $\alpha_1, \cdots, \alpha_p, \beta_1, \cdots, \beta_q$ such that:

$$\begin{aligned} m &= \alpha_1 m_1 + \cdots + \alpha_p m_p \\ n &= \beta_1 n_1 + \cdots + \beta_q n_q \end{aligned}$$

   Thus, we have:

$$\begin{aligned} v &= m + n \\ &= \alpha_1 m_1 + \cdots + \alpha_p m_p + \beta_1 n_1 + \cdots + \beta_q n_q, \end{aligned}$$

   so $\mathcal{B} \equiv \{m_1, \cdots, m_p, n_1, \cdots, n_q$ spans V.

2. To demonstrate linear independence, let scalars $\alpha_1, \cdots, \alpha_p, \beta_1, \cdots, \beta_q$ be given such that:

$$\underbrace{\alpha_1 m_1 + \cdots \alpha_p m_p}_{\in M} + \underbrace{\beta_1 n_1 + \cdots + \beta_q n_q}_{\in N} = 0_V.$$

   Since $0_V \in M \cup N$, we also have:

$$\underbrace{0_V}_{\in M} + \underbrace{0_V}_{\in N} = 0_V.$$

   But $V = M \oplus N$, so by definition the decomposition of 0_V as the sum of a vector in M and a vector in N must be unique. It follows that:

$$\begin{aligned} \alpha_1 = \cdots = \alpha_p = 0 \\ \beta_1 = \cdots = \beta_q = 0, \end{aligned}$$

   which establishes the linear independence of $\mathcal{B} \equiv \{m_1, \cdots, m_p, n_1, \cdots, n_q\}$.

" $\Leftarrow$ " For the other direction of the proof, we essentially reverse the above argument. Suppose $\beta = \{m_1, \cdots, m_p, n_1, \cdots, n_q\}$ form a basis for V. Fix an arbitrary $v \in \mathcal{V}$. Then there exists (unique) scalars $\alpha_1, \cdots, \alpha_p, \beta_1, \cdots, \beta_q$ such that:

$$v = \underbrace{\alpha_1 m_1 + \cdots \alpha_p m_p}_{\text{unique, } \in N} + \underbrace{\beta_1 n_1 + \cdots + \beta_q n_q}_{\text{unique, } \in N}$$

This establishes our desired conclusion, $V = M \oplus N$.

Remark. The above result can be extended to yield the following immediate corollary—Suppose $M_1, M_2, \cdots, M_n \leq V$, and that, for each $i = 1, \cdots, n$, the set $\mathcal{B}_i$ forms a basis for M_i. Then:

$$V = M_1 \oplus \cdots \oplus M_n$$

if and only if:

$$\mathcal{B} \equiv \mathcal{B}_1 \cup \cdots \cup \mathcal{B}_n$$

forms a basis for $\mathcal{V}$.

2.6 Lecture 5

The definitions given in the previous lecture lead naturally to the concept of orthogonal projections. For instance, suppose $b \in \mathbb{R}^2$ is a unit vector, i.e. $|b| = \langle b, b\rangle = 1$. Given any nonzero $v \in \mathbb{R}^2$ (the case $v = 0$ is trivial), define:

$$u^\star = \overline{\langle v, b\rangle} b$$

Then $v - u^\star \perp b$, since:

$$\begin{aligned}\langle v - u^\star, b\rangle &= \langle v - \overline{\langle v, b\rangle}, b\rangle \\ &= \langle v, b\rangle - \langle v, b\rangle\langle b, b\rangle = 0\end{aligned}$$

This observation is officially recorded as the theorem below.

Theorem 2.31 (**Projection Theorem**). *Let $\mathcal{V}$ be a finite-dimensional inner product space with $dim(\mathcal{V}) = n$, and let $S \leq \mathcal{V}$. Fix $v \in \mathcal{V}$ and consider the problem of finding:*

$$\inf_{s\in S} |v - s| \tag{2.2}$$

It is not always true that there exists a solution to the above equation.

1. *Suppose the optimization problem* (2.31) *is solvable, with solution $\hat{s}$. Then $\hat{s}$ is optimal if and only if:*

   $$(v - \hat{s}) \perp S$$

   Furthermore, the optimal vector $\hat{s}$ is unique.

2. *Suppose S is complete, then* (2.31) *is solvable.*

3. *Suppose S is finite-dimensional. Let $\mathcal{B} = \{b_1, \cdots, b_n\}$ be a basis for S, and define the matrix $M \in \mathbb{C}^{n\times n}$ be defined by:*

   $$M = [\langle b_i, b_j\rangle]_{n\times n}$$

   Then M is non-singular, and (2.31) *is solvable by the unique solution:*

   $$\hat{s} = \sum_{i=1}^{n} \alpha_i b_i$$

   where:

   $$\begin{bmatrix}\alpha_1 \\ \vdots \\ \alpha_n\end{bmatrix} = M^{-1}\begin{bmatrix}\langle b_1, v\rangle \\ \vdots \\ \langle b_n, v\rangle\end{bmatrix}$$

Proof.

1. " $\Rightarrow$ " Let $\hat{s}$ be optimal. Suppose by contradiction that $(v - \hat{s}) \perp S$ is false. Then there exists some $s_1 \in S$ and $\alpha \neq 0$ such that:

$$\langle s_1, v - \hat{s} \rangle = \alpha \neq 0$$

Clearly, $s_1 \neq 0_{\mathcal{V}}$. Define $\beta = \alpha / |s_1|^2 \neq 0$, and $v_{new} = \hat{s} + \beta s_1 \in S$. Then:

$$\begin{aligned} |v - s_{new}|^2 &= |v - \hat{s} - \beta s_1|^2 \\ &= |v - \hat{s}|^2 + \|\beta s_1\|^2 - \langle v - \hat{s}, \beta s_1 \rangle - \langle \beta s_1, v - \hat{s} \rangle \\ &= |v - \hat{s}|^2 + |\beta|^2 |s_1|^2 - \beta \bar{\alpha} - \alpha \bar{\beta} \\ &= |v - \hat{s}|^2 + |\alpha|^2 |s_1|^4 - 2|\alpha|^2 |s_1|^4 \\ &= |v - \hat{s}|^2 - |\alpha|^2 |s_1|^4 \\ &< |v - \hat{s}|^2 \end{aligned}$$

which indicates that $\hat{s}$ is not optimal, a contradiction.

" $\Leftarrow$ " To show the converse, let $\hat{v}$ be any vector satisfying $(v - \hat{s}) \perp S$. Then:

$$\begin{aligned} |v - s|^2 &= |(v - \hat{s}) + (\hat{s} - s)|^2 \\ &= |v - \hat{s}|^2 + |\hat{s} - s|^2 + 2\, Re(\langle v - \hat{s}, \hat{s} - s \rangle) \\ &= |v - \hat{s}|^2 + |\hat{s} - s|^2 \\ &\geq |v - \hat{s}|^2 \end{aligned}$$

since the fact that $v - \hat{s} \perp S$ and $\hat{s} - s \in S$ implies $\langle v - \hat{s}, \hat{s} - s \rangle = 0$. This shows that $\hat{s}$ is optimal.

Finally, we show that, given $v \in \mathcal{V}$, the condition $(v - \hat{s}) \perp S$ uniquely determines $\hat{s}$. Suppose there exists some $s' \in S$ such that $(v - s') \perp S$. Then:

$$\begin{aligned} |s' - \hat{s}|^2 &= \langle s' - \hat{s}, s' - \hat{s} \rangle \\ &= \langle (v - \hat{s}) - (v - s'), s' - \hat{s} \rangle \\ &= \langle v - \hat{s}, s' - \hat{s} \rangle - \langle v - s', s' - \hat{s} \rangle \\ &= 0. \end{aligned}$$

since $v - \hat{s} \perp S$, $v - s' \perp S$, and $\hat{s} - s' \in S$.

2. Define $\gamma \equiv \inf_{s \in S} |v - s|$. By definition of infimum, there exists a sequence of vectors $\{s_k\}_{k=1}^{\infty}$ in S such that:

$$\lim_{k \to \infty} |v - s_k| = \gamma$$

Let $\epsilon > 0$ be given. Then there exists some $N_0 \in \mathbb{N}$ such that $|v - s_k| < \frac{1}{2}\epsilon$ for each $k \geq N_0$. Now, let $k_1, k_2 \geq N_0$ be given. Using the triangle inequality, we have:

$$|s_{k_1} - s_{k_2}| \leq |v - s_{k_1}| + |v - s_{k_2}| < \epsilon$$

This shows that $\{s_k\}_{k=1}^{\infty}$ is a Cauchy sequence in S; since S is complete, this sequence converges, say, to some s_0, completing the proof.

3. Since $M^{\star}_{j,i} = \overline{\langle b_j, b_i \rangle} = \langle b_i, b_j \rangle = M_{i,j}$, the matrix M is Hermitian. That it is nonsingular can thus be demonstrated by showing that it is, in fact, positive definite. Let a nonzero n-tuple of scalars $a = (a_1, \cdots, a_n) \in \mathbb{C}^n \backslash \{0\}$ be arbitrarily given. Then:

$$\begin{aligned} a^{\star} M a &= \sum_{i=1}^{n} \sum_{j=1}^{n} a_i^{\star} \langle b_i, b_j \rangle a_j = \left\langle \sum_{i=1}^{n} a_i b_i, \sum_{j=1}^{n} a_j b_j \right\rangle \\ &= \left| \sum_{i=1}^{n} a_i b_i \right|^2 \geq 0 \end{aligned}$$

since the fact that $\mathcal{B}$ forms a basis for S indicates that $\sum_{i=1}^{n} a_i b_i = 0$ if and only if $a_i = 0$ for each $i = 1, \cdots, n$, contradicting the fact that a is nonzero.

Since S is finite-dimensional, it must be complete. Part b) then implies that (2.31) is solvable, while Part a) offers a method for finding the solution. In particular, let the optimal solution be of the form:

$$\hat{s} = \sum_{j=1}^{n} \alpha_j b_j$$

By Part a), $(v - \hat{s}) \perp S$, so for each $i = 1, \cdots, n$:

$$\begin{aligned} 0 = \langle b_i, v - \hat{s} \rangle &= \left\langle b_i, v - \sum_{j=1}^{n} \alpha_j b_j \right\rangle \\ &= \langle b_i, v \rangle - \sum_{j=1}^{n} \alpha_j \langle b_i, b_j \rangle \\ \Rightarrow \langle b_i, v \rangle &= \sum_{j=1}^{n} \alpha_j \langle b_i, b_j \rangle = \sum_{j=1}^{n} M_{ij} \alpha_j. \end{aligned}$$

In other words:

$$M \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} = \begin{bmatrix} \langle b_1, v \rangle \\ \vdots \\ \langle b_n, v \rangle \end{bmatrix},$$

completing the proof.

■

Orthogonal projection finds uses in filtering (de-noising), which is used to eliminate components of noise "orthogonal" to the true output of a system (e.g. Consider an output of the form $y = Ax + n$, where $n \perp R(A)$).

To simplify the form of the matrix M, we usually wish to obtain an orthonormal basis from a given basis. This is because, if the basis under consideration is orthonormal, M becomes

the identity matrix, and we thus have:

$$\begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} = \begin{bmatrix} \langle b_1, v \rangle \\ \vdots \\ \langle b_n, v \rangle \end{bmatrix}$$

The proof of the following theorem provides an algorithm for doing so.

Theorem 2.32 (**Gram-Schmidt Process for Orthonormalization**). *Each finite-dimensional inner product space has an orthonormal basis.*

Proof. This theorem can be proved by directly constructing an orthonormal basis for an arbitrarily given finite-dimensional inner product space $\mathcal{V}$. If $\mathcal{V} = \{\mathbf{0}\}$, then the empty set ϕ is an orthonormal basis for $\mathcal{V}$. Suppose $\mathcal{V} \neq \{\mathbf{0}\}$, and $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ is a basis for $\mathcal{V}$. Define subspaces $\mathcal{W}_j$ of $\mathcal{V}$, for each $i = 1, 2, \cdots, n$ as:

$$\mathcal{W}_j = \text{ span}\{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_j\}$$

Evidently, $\mathcal{W}_1 \subset \mathcal{W}_2 \subset \cdots \subset \mathcal{W}_n \subset \mathcal{V}$.

To construct a basis for $\mathcal{W}_1 = \text{span}\{\mathbf{v}_1\}$, let:

$$\mathbf{u}_1 = \frac{\mathbf{v}_1}{\|\mathbf{v}_1\|}$$

Since $\mathbf{u}_1$ is a unit vector, $\{\mathbf{u}_1\}$ is an orthonormal basis for $\mathcal{W}$. Suppose an orthonormal basis $\{\mathbf{u}_1, \cdots \mathbf{u}_k\}$ for $\mathcal{W}_k$ has been defined, for some $k = 1, 2, \cdots, n-1$. Define $\mathbf{P}_{\mathcal{W}_k}$ as the orthogonal projection onto $\mathcal{W}_k$, and let:

$$\mathbf{u}_{k+1} = \frac{\mathbf{v}_{k+1} - \mathbf{P}_{\mathcal{W}_k}(\mathbf{v}_{k+1})}{\|\mathbf{v}_{k+1} - \mathbf{P}_{\mathcal{W}_k}(\mathbf{v}_{k+1})\|}$$

By induction, an orthonormal basis $\{\mathbf{u}_1, \cdots, \mathbf{u}_n\}$ for $\mathcal{W}_n = \mathcal{V}$ can be found. ■

A numerical example of the Gram-Schmidt orthonormalization process is given below.

Example. Consider the following set in $\mathbb{R}^3$:

$$\mathcal{B} = \left\{ \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix}, \begin{bmatrix} 3 \\ 4 \\ 5 \end{bmatrix} \right\} \equiv \{v_1, v_2, v_3\}$$

Suppose we wish to obtain an orthonormal basis $\mathcal{B}' = \{v_1', v_2', v_3'\}$ with the same span as $\mathcal{B}$. First, note that $v_1 - 2v_2 + v_3 = 0$, so it suffices to restrict our attention to $\{v_1, v_2\}$.

Applying the Gram-Schmidt orthonormalization procedure, we have:

$$v_1' \propto \frac{v_1}{|v_1|} = \frac{1}{\sqrt{14}} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

$$v_2' \propto v_2 - \frac{\langle v_2, v_1 \rangle}{\langle v_1, v_1 \rangle} v_1 = \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix} - \frac{20}{14} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = \frac{1}{7} \begin{bmatrix} 4 \\ 1 \\ -2 \end{bmatrix}$$

Normalizing the above orthogonal vectors, we have:

$$v_1' = \frac{1}{\sqrt{14}} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

$$v_2' = \frac{1}{\sqrt{21}} \begin{bmatrix} 4 \\ 1 \\ -2 \end{bmatrix}$$

2.7 Lecture 6

Definition 2.33 (**Self-Adjoint Maps**). *Given an inner product space $(H, \mathbb{F}, \langle \cdot, \cdot \rangle_H)$, let $A : H \to H$ be a continuous linear map with adjoint $A^\star : H \to H$. The map A is said to be **self-adjoint** if $A = A^\star$, i.e. for each $x, y \in H$:*

$$\langle x, Ay \rangle_H = \langle Ax, y \rangle_H$$

Example (**Hermitian Matrices**). Let the linear map $A : \mathbb{F}^n \to \mathbb{F}^n$ be represented by a matrix $A = (a_{ij})_{i,j \in \{1, \cdots, m\}} \in \mathbb{F}^{n \times n}$. Then A is self-adjoint if and only if the matrix A is Hermitian. Equivalently, $A = A^\star$, meaning $a_{ij} = \overline{a_{ji}}$ for each $i, j \in \{1, \cdots, n\}$, or that A is equal to its complex conjugate transpose.

Definition 2.34 (**Unitary Matrix, Orthogonal Matrix**). *A matrix $U \in \mathbb{F}^{n \times n}$ is called a **unitary matrix** if:*

$$U^\star = U^{-1}.$$

*Equivalently, U is a unitary matrix if and only if the n columsn and n rows of U form orthonormal bases for $\mathbb{F}^n$. If $\mathbb{F} = \mathbb{R}$, such a matrix is said to be **orthogonal**.*

Below, we present an important theorem often used in mathematical optimization.

Theorem 2.35. *Let $\mathcal{V}$ and $\mathcal{U}$ be finite-dimensional inner product spaces over $\mathbb{C}$ or $\mathbb{R}$, with dimensions n and m, respectively. Let $\boldsymbol{L} \in \mathcal{L}(\mathcal{U}, \mathcal{V})$ be given, and define $r = \text{rank}(\mathbf{L})$. Then there exist ordered, orthonormal bases $\mathcal{B}_\mathcal{U} = \{\mathbf{u}_1, \cdots, \mathbf{u}_r, \mathbf{u}_{r+1}, \cdots, \mathbf{u}_n\}$, for $\mathcal{U}$, and $\mathcal{B}_\mathcal{V} = \{\mathbf{v}_1, \cdots, \mathbf{v}_r, \mathbf{v}_{r+1}, \cdots, \mathbf{v}_m\}$, for $\mathcal{V}$, such that:*

1. *$\{\mathbf{u}_1, \cdots, \mathbf{u}_r\}$ is an ordered, orthonormal basis for $\mathcal{K}er(\mathbf{L})^\perp = \mathcal{I}m(\mathbf{L}^\dagger)$.*
2. *$\{\mathbf{u}_{r+1}, \cdots, \mathbf{u}_m\}$ is an ordered, orthonormal basis for $\mathcal{K}er(\mathbf{L}) = \mathcal{I}m(\mathbf{L}^\dagger)^\perp$.*
3. *$\{\mathbf{v}_1, \cdots, \mathbf{v}_r\}$ is an ordered, orthonormal basis for $\mathcal{I}m(\mathbf{L}) = \mathcal{K}er(\mathbf{L}^\dagger)^\perp$.*
4. *$\{\mathbf{v}_{r+1}, \cdots, \mathbf{v}_m\}$ is an ordered, orthonormal basis for $\mathcal{I}m(\mathbf{L})^\perp = \mathcal{K}er(\mathbf{L}^\dagger)$.*
5. *$\mathbf{L}\mathbf{u}_i = \sigma_i \mathbf{v}_i$, $\mathbf{L}\mathbf{v}_i = \sigma_i \mathbf{u}_i$, where $\sigma_i \geq 0$, for each $i = 1, \cdots, r$.*

*In particular, $\sigma_1, \cdots, \sigma_r$ are said to be the **singular values** of $\mathbf{L}$, $\mathbf{u}_1, \cdots, \mathbf{u}_r$ are said to be the **right singular vectors** of $\mathbf{L}$, while $\mathbf{v}_1, \cdots, \mathbf{v}_r$ are said to be the **left singular vectors** of $\mathbf{L}$.*

Proof. To demonstrate Part 2 of this theorem, consider the positive operator $\mathbf{L}^\dagger \mathbf{L} \in \mathcal{L}(\mathcal{U})$. Since $r = \text{rank}(\mathbf{L}) = \text{rank}(\mathbf{L}^\dagger \mathbf{L})$, there exists an ordered, orthonormal basis $\mathcal{B}_\mathcal{U} = \{\mathbf{u}_1, \cdots, \mathbf{u}_r, \mathbf{u}_{r+1}, \cdots, \mathbf{u}_n\}$ consisting of eigenvectors of $\mathbf{L}^\dagger \mathbf{L}$, with corresponding eigenvalues satisfying:

$$\lambda_1 \geq \cdots \geq \lambda_r > 0 = \lambda_{r+1} = \cdots + \lambda_n.$$

For each $i = r + 1, \cdots, n$:

$$\|\mathbf{L}\mathbf{u}_i\|^2 = \langle \mathbf{L}\mathbf{u}_i, \mathbf{L}\mathbf{u}_i \rangle = \langle \mathbf{L}^\dagger \mathbf{L}\mathbf{u}_i, \mathbf{u}_i \rangle = \lambda_i \|\mathbf{u}_i\|^2 = 0,$$

so $\mathbf{L}\mathbf{u}_i = \mathbf{0}$. Thus, $\text{span}(\{\mathbf{u}_{r+1}, \cdots, \mathbf{u}_n\}) \subset \mathcal{K}er(\mathbf{L})$. Conversely, suppose $\mathbf{v} = \sum_{i=1}^n a_u \mathbf{u}_i \in \mathcal{K}er(\mathbf{L})$. Then:

$$\begin{aligned} 0 = \langle \mathbf{L}\mathbf{v}, \mathbf{L}\mathbf{v}\rangle &= \left\langle \sum_{i=1}^n a_i \mathbf{L}\mathbf{u}_i, \sum_{j=1}^n a_j \mathbf{L}\mathbf{u}_j \right\rangle = \sum_{i=1}^n \sum_{j=1}^n a_i a_j^\star \langle \mathbf{L}\mathbf{u}_i, \mathbf{L}\mathbf{u}_j\rangle \\ &= \sum_{i=1}^n \sum_{j=1}^n a_i a_j^\star \langle \mathbf{L}^\dagger \mathbf{L}\mathbf{u}_i, \mathbf{u}_j\rangle = \sum_{i=1}^n \sum_{j=1}^n a_i a_j^\star \lambda_i \langle \mathbf{u}_i, \mathbf{u}_j\rangle = \sum_{i=1}^n \sum_{j=1}^n a_i a_j^\star \lambda_i \delta_{ij} \\ &= \sum_{i=1}^r |a_i|^2 \lambda_i^2, \end{aligned}$$

Thus,$a_i = 0$ for each $i = 1, 2, \cdots, r$, so $\mathbf{v} \in \text{span}(\{\mathbf{u}_{r+1}, \cdots, \mathbf{u}_n\})$, which implies $\mathcal{K}er(\mathbf{L}) \subset \text{span}(\{\mathbf{u}_{r+1}, \cdots, \mathbf{u}_n\})$. This establishes Part 2, i.e. $\mathcal{K}er(\mathbf{L}) \subset \text{span}(\{\mathbf{u}_{r+1}, \cdots, \mathbf{u}_n\})$. Since $\{\mathbf{u}_1, \cdots, \mathbf{u}_r, \mathbf{u}_{r+1}, \cdots, \mathbf{u}_n\}$ is orthonormal, $\{\mathbf{u}_1, \cdots, \mathbf{u}_r\}$ is thus an ordered, orthonormal basis for $\mathcal{I}m(\mathbf{L}) = \mathcal{K}er(\mathbf{L}^\dagger)$, which confirms Part 1.

To prove Part 5, define $\sigma_i = \sqrt{\lambda_i}$ for each $i = 1, 2, \cdots, n$. Then:

$$\mathbf{L}^\dagger \mathbf{L}\mathbf{u}_i = \lambda \mathbf{u}_i = \sigma_i^2 \mathbf{u}_i$$

In particular, $\sigma_i = 0$ for each $i = r+1, \cdots, n$. *Define* $\mathbf{v}_i = (1/\sigma_i)\mathbf{L}\mathbf{u}_i$. Then:

$$\mathbf{L}\mathbf{u}_i = \begin{cases} \sigma_i \mathbf{v}_i, & i = 1, 2, \cdots r, \\ 0, & i \leq r, \end{cases}$$

$$\mathbf{L}^\dagger \mathbf{v}_i = \begin{cases} \sigma_i \mathbf{u}_i, & i = 1, 2, \cdots r, \\ 0, & i \leq r. \end{cases}$$

This establishes Part 5.

To prove Part 3, for each $i, j = 1, \cdots, r$:

$$\langle \mathbf{v}_i, \mathbf{v}_j\rangle = \frac{1}{\sigma_i \sigma_j}\langle \mathbf{L}\mathbf{u}_i, \mathbf{L}\mathbf{u}_j\rangle = \frac{1}{\sigma_i \sigma_j}\langle \mathbf{L}^\dagger \mathbf{L}\mathbf{u}_i, \mathbf{u}_j\rangle = \frac{\sigma_i}{\sigma_j} = \delta_{ij},$$

so $\{\mathbf{v}_1, \cdots, \mathbf{v}_r\}$ is an ordered, orthonormal basis for $\mathcal{I}m(\mathbf{L}) = \mathcal{K}er(\mathbf{L}^\dagger)^\perp$. Using the Extension Theorem (Alternate Proof for Theorem 2.84) and the Gram-Schmidt Process (Theorem 6.31), choose normalized vectors $\mathbf{v}_{r+1}, \cdots, \mathbf{v}_n$ such that $\{\mathbf{v}_1, \cdots, \mathbf{v}_r, \cdots, \mathbf{v}_{r+1}, \cdots, \mathbf{v}_m\}$ such that $\mathcal{B}_\mathcal{V}$ is an ordered, orthonormal basis for $\mathcal{V}$. Evidently, $\{\mathbf{v}_{r+1}, \cdots, \mathbf{v}_m\}$ is an ordered, orthonormal basis for $\mathcal{K}er(\mathbf{L}^\dagger)$, which verifies Part 4. It is not difficult to observe that:

$$\mathbf{L}\mathbf{L}^\dagger \mathbf{v}_i = \sigma_i \mathbf{L}\mathbf{u}_i = \sigma_i^2 \mathbf{v}_i,$$

so $\mathbf{v}_1, \cdots, \mathbf{v}_r$ are eigenvalues of $\mathbf{L}\mathbf{L}^\dagger$ corresponding to $\lambda_1 (= \sigma_1^2), \cdots, \lambda_r (= \sigma_r^2)$. ■

The singular value decomposition can also be presented as shown below.

Theorem 2.36 (**Singular Value Decomposition for Matrices**). *Let $M \in \mathbb{C}^{m\times n}$ with rank(M) = r. Then there exist unitary matrices $U \in \mathbb{C}^{m\times m}$ and $V \in \mathbb{C}^{n\times n}$ such that:*

$$M = U\Sigma V = U \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} V^\star$$

where:

$$\Sigma_1 = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix},$$

*where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ are called the **singular values** of M, and the representation above is called the **singular-value decomposition** of M.*

Theorem 2.37. *Let $M \in \mathbb{C}^{m\times n}$ with rank(M) = r, and let $M = U\Sigma V^\star$ be the singular-value decomposition of M. Partition U and V as:*

$$U = \begin{bmatrix} U_1 & U_2 \end{bmatrix}, \qquad V = \begin{bmatrix} V_1, V_2 \end{bmatrix},$$

where $U_1, V_1 \in \mathbb{C}^{r\times r}$. Then:

1. *The columns of U_1 and U_2 form orthonormal bases for $R(M)$ and $N(M^\star)$, respectively. (Note—these are subspaces of $\mathbb{C}^m$.*
2. *The columns of V_1 and V_2 form orthonormal bases for $R(M)^\star$ and $N(M)$, respectively. (Note—these are subspaces of $\mathbb{C}^n$.*

Chapter 3

Dynamical Systems

3.1 Lecture 7

In this section, we investigate properties of differential equations. In particular, given a differential equation:

$$\dot{x} = f(x,t), \quad x(t_0) = x_0$$

where $x(t) \in \mathbb{R}^n$, $f(x,t) : \mathbb{R}^n \times \overline{\mathbb{R}^+} \to \mathbb{R}^n$, we are interested in the conditions under which:

1. A solution exists, i.e. there exists some $x(t)$, defined for all $t \geq t_0$, satisfying the given differential equation.
2. The solution is unique.

Definition 3.1 (**Piecewise Continuity**). *The function $f(x,t) : \mathbb{R}^n \times \overline{\mathbb{R}^+} \to \mathbb{R}^n$ is said to be **piecewise continuous in** t if, in any compact interval, $f(x,\cdot) : \overline{\mathbb{R}^+} \to \mathbb{R}^n$ is continuous except at a finite number of points.*

Definition 3.2 (**Lipschitz Continuity**). *The function $f(x,t) : \mathbb{R}^n \times \overline{\mathbb{R}^+} \to \mathbb{R}^n$ is said to be **Lipschitz continuous in** x if, for each t, there exists a piecewise continuous function $\kappa(\cdot) : \overline{\mathbb{R}^+} \to \overline{\mathbb{R}^+}$ such that:*

$$|f(x,t) - f(y,t)| \leq \kappa(t) \cdot |x - y|$$

*for each $x, y \in \mathbb{R}^n$, and $t \in \overline{\mathbb{R}^+}$. This inequality is called the **Lipschitz condition**.*

We wish to present a theorem for the existence of a unique solution to a differential equation with certain continuity restrictions. However, before presenting the theorem, we first examine the following lemma.

Lemma 3.3 (**Bellman-Gronwall Lemma**). *Let $u(\cdot), k(\cdot)$ be real-valued, piecewise continuous functions on $\overline{\mathbb{R}^+}$, and assume that $u(t), k(t) > 0$ on $\overline{\mathbb{R}^+}$. If, for some differentiable, non-decreasing function $c(t)$:*

$$u(t) \leq c(t) + \int_{t_0}^{t} k(\tau)u(\tau)\,d\tau \equiv Z(t) \tag{3.1}$$

then:

$$u(t) \leq c(t)e^{\int_{t_0}^{t} k(\tau)\,d\tau}$$

Proof. Without loss of generality, suppose $t > t_0$. Define:

$$Z(t) = c(t) + \int_{t_0}^{t} k(\tau)u(\tau)\,d\tau,$$

then $u(t) \leq Z(t)$. In differential form:

$$\begin{aligned} \frac{d}{dt}Z(t) &= c'(t) + k(t)u(t) \\ Z(t_0) &= c(t_0) \end{aligned}$$

Multiplying both sides of (3.1) by the non-negative function:

$$k(t)e^{-\int_{t_0}^{t} k(\tau)\,d\tau} \geq 0$$

we find:

$$\begin{aligned} 0 &\geq \left[u(t) - Z(t)\right] \cdot k(t)e^{-\int_{t_0}^{t} k(\tau)\,d\tau} \\ &\geq \left(\frac{d}{dt}Z(t) - c'(t) - Z(t)k(t)\right) \cdot e^{-\int_{t_0}^{t} k(\tau)\,d\tau} \\ &= \left(\frac{d}{dt}Z(t) - Z(t)k(t)\right) \cdot e^{-\int_{t_0}^{t} k(\tau)\,d\tau} - c'(t) \underbrace{e^{-\int_{t_0}^{t} k(\tau)\,d\tau}}_{\leq 1, \text{ since } k(t)\geq 0} \\ &\geq \left(\frac{d}{dt}Z(t) - Z(t)k(t)\right) \cdot e^{-\int_{t_0}^{t} k(\tau)\,d\tau} - c'(t) \\ &= \frac{d}{dt}\left(Z(t) \cdot e^{-\int_{t_0}^{t} k(\tau)\,d\tau} - c(t)\right) \end{aligned}$$

Thus, the function $Z(t) \cdot e^{-\int_{t_0}^{t} k(\tau)\,d\tau} - c(t)$ is decreasing, and must at any time t be less than its value at t_0:

$$\begin{aligned} Z(t) \cdot e^{-\int_{t_0}^{t} k(\tau)\,d\tau} - c(t) &\leq Z(t_0) - c(t_0) = 0 \\ \Rightarrow u(t) \leq Z(t) &\leq c(t) \cdot e^{\int_{t_0}^{t} k(\tau)\,d\tau} \end{aligned}$$

■

Theorem 3.4 (**Fundamental Theorem of Differential Equations**). *Consider the differential equation:*

$$\dot{x} = f(x,t), \quad x(t_0) = x_0$$

where $f(x,t) : \mathbb{R}^n \times R_+ \to \mathbb{R}^n$ is piecewise continuous in t and Lipschitz continuous in x. Then there exists a unique *function of time $\phi(\cdot) : \overline{\mathbb{R}^+} \to \mathbb{R}^n$ that is continuously differentiable almost everywhere, and satisfies:*

$$\begin{aligned} \phi(t_0) &= x_0 \\ \dot{\phi}(t,0) &= f(\phi(t),t), \end{aligned}$$

for each $t \in [t_0, t_1]\backslash D$, where D denotes the set of discontinuity points of f as a function of t.

Proof. Construct a sequence of continuous functions:

$$x_{m+1}(t) \equiv x_0 + \int_{t_0}^{t} f(x_m(\tau), \tau) d\tau, \quad m = 0, 1, 2, \cdots$$

where $x_0(t_0) = t_0$. The idea is to show that

1. The sequence of continuous functions:

   $$\{x_m(\cdot)\}_0^\infty$$

   converges to: a continuous function $\phi(\cdot) : \overline{\mathbb{R}^+} \to \mathbb{R}^n$,

2. $\phi(\cdot)$ is a solution of the given ODE, i.e.:

   $$\dot{\phi} = f(\phi, t), \quad \phi(t_0) = x_0.$$

3. $\phi(\cdot)$ is the unique solution. This technique is known as the "construction of a solution by iteration."

We proceed to prove each of the above claims.

1. To show that $\phi(\cdot)$ is a continuous function, we first demonstrate that $\{x_m(\cdot)\}_0^\infty$ is a *Cauchy sequence* in the *Banach space* $(C([t_1, t_2], \mathbb{R}^n), \mathbb{R}, \|\cdot\|_\infty)$, where $t_0, t \in [t_1, t_2]$:

$$\begin{aligned}
\|x_{m+1}(t) - x_m(t)| &\equiv \left\| \int_{t_0}^{t} \left[f(x_m(\tau), \tau) - f(x_{m-1}(\tau), \tau) \right] d\tau \right\| \\
&\leq \int_{t_0}^{t} \|f(x_m(\tau), \tau) - f(x_{m-1}(\tau), \tau)\| \, d\tau \\
&\leq \int_{t_0}^{t} \kappa(\tau) \cdot \|x_m(\tau) - x_{m-1}(\tau)\| \, d\tau \\
&\leq \bar{\kappa} \cdot \int_{t_0}^{t} \|x_m(\tau) - x_{m-1}(\tau)\| \, d\tau
\end{aligned}$$

where $\kappa(t)$ is a piecewise continuous function arising from the fact that $f(x, t)$ is Lipschitz continuous in x, and where we have defined $\bar{\kappa} = \sup_{t \in [t_1, t_2]} \kappa(t)$. By the definition of $\{x_m(\cdot)\}_0^\infty$:

$$\begin{aligned}
x_1(t) &\equiv x_0 + \int_{t_0}^{t} f(x_0, \tau) \, d\tau, \quad t \in [t_1, t_2] \\
\therefore \|x_1(t) - x_0\| &\leq \int_{t_0}^{t} \|f(x_0, \tau)\| \, d\tau \leq \int_{t_1}^{t_2} \|f(x_0, \tau)\| \, d\tau \equiv M
\end{aligned}$$

Since x_0 is specified, M is known.

Thus, if we define $T = |t - t_0|$ and apply the recursive bound derived above, we have:

$$\begin{aligned}
\|x_2(t) - x_1(t)\| &\leq M\overline{\kappa}|t - t_0| = M\overline{\kappa}T \\
\|x_3(t) - x_2(t)\| &\leq \frac{1}{2}M\overline{\kappa}^2T^2 \\
&\vdots \\
\|x_2(t) - x_1(t)\| &\leq \frac{1}{m!}M\overline{\kappa}^mT^m
\end{aligned}$$

Next, recall that $\|f(\cdot)\|_\infty = \max\{|f(t)|, t \in [t_1, t_2]\}$. To see that $\{x_m(\cdot)\}_{m=0}^\infty$ is a Cauchy sequence in $(\mathscr{C}([t_1, t_2], \mathbb{R}^n), \mathbb{R}, \|\cdot\|_\infty)$, observe that for any $m, p \in \mathbb{N}$, we have:

$$\begin{aligned}
\|x_{m+p}(\cdot) - x_m(\cdot)\|_\infty &= \left\| \sum_{k=0}^{p-1} \left[x_{m+k+1}(\cdot) - x_{m+k}(\cdot) \right] \right\|_\infty \\
&\leq \sum_{k=0}^{p-1} \left\| \left[x_{m+k+1}(\cdot) - x_{m+k}(\cdot) \right] \right\|_\infty \\
&\leq M \cdot \sum_{k=0}^{p-1} \frac{(\overline{\kappa}T)^{m+k}}{(m+k)!} \\
&\leq M \frac{(\overline{\kappa}T)^m}{m!} \cdot \sum_{k=0}^{p-1} \frac{(\kappa T)^k}{k!} \\
&\leq M \frac{(\overline{\kappa}T)^m}{m!} e^{\overline{\kappa}T}
\end{aligned}$$

Thus, given any $\epsilon > 0$, we can always choose a sufficiently large $m \in \mathbb{N}$ such that, for each $p \in \mathbb{N}$:

$$\|x_{m+p}(\cdot) - x_m(\cdot)\|_\infty < \epsilon$$

By definition, $\{x_m(\cdot)\}_{m=0}^\infty$ is Cauchy.

2. Next, we must show that $\phi(\cdot)$ is a solution of the given differential equation, i.e. $\dot{\phi} = f(\phi, t), \phi(t_0) = x_0$. It is sufficient to show that:

$$\phi(t) = x_0 + \int_{t_0}^t f(\phi(\tau), \tau) d\tau,$$

since the differential equation would follow by differentiating with respect to t on both sides, while the initial condition can be obtained substituting $t = t_0$ on both sides.

By construction, we already have:

$$x_{m+1}(t) = x_0 + \int_{t_0}^t f(x_m(\tau), \tau) d\tau$$

Moreover, by the proof in Part 1, $x_m(\cdot) \to \phi(\cdot)$ on $[t_1, t_2]$ as $m \to \infty$. It remains to verify that:

$$\int_{t_0}^{t} f(x_m(\tau), \tau) d\tau \longrightarrow \int_{t_0}^{t} f(\phi(\tau), \tau) d\tau$$

as $m \to \infty$. This can be done straightforwardly, by once again observing the convergence of $\{x_m(t)\}_{m=1}^{\infty}$ (with respect to the infinity norm) derived in (1):

$$\begin{aligned}
&\left| \int_{t_0}^{t} \left[f(x_m(\tau), \tau) - f(\phi(\tau), \tau) \right] d\tau \right| \\
=& \int_{t_0}^{t} \left| f(x_m(\tau), \tau) - f(\phi(\tau), \tau) \right| d\tau \\
\leq& \int_{t_0}^{t} \kappa(\tau) \cdot |x_m(\tau) - \phi(\tau)| \, d\tau \\
\leq& \bar{\kappa} \cdot \|x_m(\cdot) - \phi(\cdot)\|_{\infty} \cdot T \\
\leq& \kappa \cdot M e^{\kappa T} \cdot \frac{(\kappa T)^m}{m!} \cdot T,
\end{aligned}$$

which approaches 0 as $m \to \infty$.

3. Finally, uniqueness can be demonstrated using the Bellman-Gronwell Lemma. Suppose $\psi(t)$ is a solution satisfying:

$$\dot{\psi(t)} = f(\psi(t), t), \quad \psi(t_0) = x_0$$

Then:

$$\begin{aligned}
|\psi(t) - \phi(t)| &= \left| \int_{t_0}^{t} \left[f(\psi(\tau), \tau)) - f(\phi(\tau), \tau)) \right] d\tau \right| \\
&\leq \int_{t_0}^{t} |f(\psi(\tau), \tau) - f(\phi(\tau), \tau)| \, d\tau \\
&\leq \int_{t_0}^{t} \kappa(t) \cdot |\psi(\tau) - \phi(\tau)| \, d\tau
\end{aligned}$$

By the Bellman-Gronwell Lemma (here, $c_1 = 0$), we have $|\psi(t) - \phi(t)| = 0$, i.e. $\psi(t) = \phi(t)$ for each $t \in [t_1, t_2]$.

■

Example. Consider the time-variant system:

$$\begin{cases} \dot{x}(t) &= A(t)x(t) + B(t)u(t), \\ x(t_0) &= x_0 \end{cases}$$

Show that the solution to this ODE is unique.

Solution :

Suppose $\phi(t)$ and $\psi(t)$ are two solutions to the given ODE. Then $\phi(t_0) = \psi(t_0) = x_0$, and:

$$\dot{\phi}(t) = A(t)\phi(t) + B(t)u(t)$$
$$\dot{\psi}(t) = A(t)\psi(t) + B(t)u(t)$$

Then:

$$\begin{aligned} |\psi(t) - \phi(t)| &= \left| \int_{t_0}^{t} \left[\dot{\psi}(t) - \dot{\phi}(t)\right] d\tau \right| \\ &\leq \int_{t_0}^{t} |A(\tau)| \cdot |\psi(\tau) - \phi(\tau)| \, d\tau \\ &\leq \|A(t)\|_\infty \cdot \int_{t_0}^{t} |\psi(\tau) - \phi(\tau)| \, d\tau \end{aligned}$$

where the infinity norm of $A(t)$ was taken over the interval $[t_0, t_1]$. By the Bellman-Gronwell Lemma, $|\psi(t) - \phi(t)| = 0$, so $\psi(t) = \phi(t)$ for each $t \geq 0$. The proof is done.

Example (**Reverse-Time Differential Equation**). Consider again the differential equation:

$$\dot{x} = f(x, t), \qquad x(t_0) = x_0$$

Suppose $f(x, t)$ satisfies the hypotheses of the Fundamental Theorem, so that the solution exists and is unique for $t \geq t_0$.

Now, consider $\tau \in (0, t_0)$, i.e. $\tau = t_0 - t$ for some $t \in (0, t_0)$. Show that there exists a trajectory $z(\tau)$ such that $z(\tau) = x(t)$.

Solution :

The proof can be done by constructing a differential equation for $z(\tau)$ that satisfies the constraints imposed in the Fundamental Theorem:

$$\begin{aligned} \frac{d}{d\tau} z(\tau) &= -\frac{d}{dt} x(t) = -f(x(t), t) = -f(z(\tau), t_0 - \tau) \\ &\equiv \overline{f}(z(\tau), \tau) \end{aligned}$$

Since $f(x, t)$ is piecewise continuous in t and Lipschitz continuous in x, we see that $\overline{f}(z, \tau)$ is piecewise continuous in τ and Lipschitz continuous in z. This is because, given any $z_1(\tau), z_2(\tau)$, if we define $\hat{z}_1(\tau) \equiv z_1(t_0 - \tau) = z_1(t)$ and $\hat{z}_2(\tau) \equiv z_2(t_0 - \tau) = z_2(t)$:

$$\begin{aligned} \|\overline{f}(z_2(\tau), \tau) - \overline{f}(z_1(\tau), \tau)\| &= \|f(z_2(\tau), t_0 - \tau) - f(z_1(\tau), t_0 - \tau)\| \\ &= \|f(\hat{z}_2(t), t) - f(\hat{z}_1(t), t)\| \end{aligned}$$

Thus, given the relation between f and $\overline{f}$, $f(x, t)$ is Lipschitz in x if and only if $\overline{f}(z, \tau)$ is Lipschitz in z. The hypotheses of the Fundamental Theorem are thus satisfied, so $\overline{f}$ is Lipschitz in z.

3.2 Lecture 7 Discussion

Definition 3.5 (**Locally Lipschitz**). *A system of differential equations is said to be* ***locally Lipschitz*** *in $\mathbb{R}^n$ if, for each $r > 0$, there exists some $L_r > 0$ such that:*

$$|f(x) - f(y)| \leq L_r \cdot |x - y|$$

for each $x, y \in B_r(0)$, where $B_r(0)$ is the n-dimensional ball of radius r centered at 0.

Theorem 3.6 (**Mean-Value Theorem in $\mathbb{R}^n$**). *Suppose S is an open subset of $\mathbb{R}^n$, and assume that $f : S \to \mathbb{R}^m$ is differentiable at each point of S. Let $x, y \in S$ be given such that $\lambda x + (1 - \lambda)y \in S$ for each $\lambda \in [0, 1]$. Then there exists some $\lambda \in (0, 1)$ such that:*

$$|f(y) - f(x)| \leq |f'(z)| \cdot |y - x|$$

where $z = \lambda x + (1 - \lambda)y$. Here, $f'(z)$ denotes the Jacobian of f at z.

Note. The above version of the Mean-Value Theorem follows directly from Apostol [1], Theorem 12.9, pg. 355.

Example (**Discussion 4, Problem 1**). Consider the following system of differential equations.

$$\begin{aligned} \dot{x}_1 &= x_1^2 + x_2^2 \\ \dot{x}_2 &= x_1^2 - x_2^2 \end{aligned}$$

As in Lecture 7, $f(x_1, x_2) = (\dot{x}_1, \dot{x}_2) = (x_1^2 + x_2^2, x_1^2 - x_2^2)$. Prove that f is locally Lipschitz but not (globally) Lipschitz.

Solution :

1. To show that the system is locally Lipschitz, we apply the n-dimensional Mean-Value Theorem. Fix $r > 0$, and let $x, y \in B_r(0)$ be given. Since f is continuously differentiable throughout $\mathbb{R}^2$, we can calculate its Jacobian at any point $(z_1, z_2) \in B_r(0)$ as:

$$f'(z_1, z_2) = \left\| \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} \end{bmatrix}_{(z_1, z_2)} \right\|_F = \left\| \begin{bmatrix} 2z_1 & 2z_2 \\ 2z_1 & -2z_2 \end{bmatrix} \right\|_F < \sqrt{8}\, r$$

   where the subscript F denotes the Frobenius norm. (Since all norms are equivalent on $\mathbb{R}^2$, norms can be arbitrarily chosen so long as their use is consistent). Note that, since $(z_1, z_2) \in B_r(0)$, we have $z_1^2 + z_1^2 < r$. Since this holds for any $(z_1, z_2) \in B_r(0)$, which is convex, we can apply the Mean-Value Theorem to get:

$$|f(x) - f(y)| \leq 8r \cdot |x - y|$$

   as desired.

2. Suppose by contradiction that there exists some $L > 0$ such that:

$$|f(x) - f(y)|_1 \leq L \cdot |x - y|_1$$

where the subscript "∞" denotes the one-norm, defined on $\mathbb{R}^2$ by:

$$\left| \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \right| = |v_1| + |v_2|$$

The local Lipschitz property of f, as demonstrated above, implies that that this inequality may in fact hold if x, y are constrained to be within a sufficiently small ball centered at the origin $(0, 0)$. Thus, to achieve a contradiction, we must take $x, y \in \mathbb{R}^2$ to be sufficiently large, e.g.:

$$x \equiv \begin{bmatrix} 2L \\ 0 \end{bmatrix}, \qquad y = \begin{bmatrix} L \\ 0 \end{bmatrix}$$

In this case:

$$|f(x) - f(y)|_1 = \left| \begin{bmatrix} 4L^2 \\ 0 \end{bmatrix} - \begin{bmatrix} L^2 \\ 0 \end{bmatrix} \right| = 3L^2$$
$$L \cdot |x - y|_1 = L \cdot \left| \begin{bmatrix} 2L \\ 0 \end{bmatrix} - \begin{bmatrix} L \\ 0 \end{bmatrix} \right| = L^2$$

Clearly, $|f(x) - f(y)|_1 \leq L \cdot |x - y|_1$ does *not* hold in this case, a contradiction.

Example (**Discussion 4, Problem 2**). Consider the following linear system:

$$\dot{x} = A(t)x(t) + B(t)u(t)$$
$$x(t_0) = x_0$$

Provide a sufficient condition for the linear system to have a unique solution.

Solution:

By the Fundamental Theorem, f should be:

- Piecewise continuous in t, for any given x, and
- Lipschitz continuous in x, for any given t, with the Lipschitz constant piecewise continuous in time.

Thus, if we:

1. Fix x —$A(t), B(t), u(t)$ are piecewise continuous in t.
2. Fix t —Observe that:

$$|f(x_1, t) - f(x_2, t)| \leq \|A(t)\|_i \cdot |x_1(t) - x_2(t)|$$

Thus, it suffices to show that $A(t)$ is bounded and piecewise continuous.

Example. Consider the following linear system:

$$\dot{x} = Ax(t)$$
$$x(0) = x_0$$

where $A \in \mathbb{R}^{n\times n}$ is nonzero. Now, suppose the initial state x_0 undergoes a variation $\tilde{x}_0$ to become $\hat{x}(0) = x_0 + \tilde{x}_0$. Then the system evolves according to:

$$\dot{\hat{x}} = Ax(t)$$
$$\hat{x}(0) = x_0$$

By subtracting each corresponding equation in the above system, we arrive at a third system for the error of the system:

$$\dot{\tilde{x}} = A\tilde{x}(t)$$
$$\tilde{x}(0) = \tilde{x}_0$$

Use the Bellman-Gronwell Lemma to show that, as $|\tilde{x}_0| \to 0$, we have $|\tilde{x}(t)| \to 0$ for any $t \in [0, T]$.

Solution:

The problem can be solved by appropriately bounding the error of the state at time t. Rewriting the system for $\tilde{x}(t)$ as an integral equation, we have:

$$\tilde{x}(t) = \tilde{x}_0 + \int_0^t A\tilde{x}(t)\,dt$$
$$\Rightarrow |\tilde{x}(t)| \le |\tilde{x}_0| + \int_0^\infty \|A\| \cdot |\tilde{x}(t)|\,dt$$

where the Cauchy-Schwarz Inequality has been applied. Since $|\tilde{x}_0|$ is constant and $\|A\| > 0$, we can apply Bellman-Gronwell lemma to find:

$$|\tilde{x}(t)| \le |\tilde{x}_0| \cdot e^{\int_0^\infty \|A\|\,dt} = |\tilde{x}_0| \cdot e^{\|A\|\,t}$$

Remark. Essentially, the problem statement claims that the error in the state $x(t)$ at any given time t can be made arbitrarily small by adequately reducing the error in the initial state x_0. The solution reveals that although this is true, the bound increases exponentially with t.

Example (**Differential Version**). Let $x(t)$ be a non-negative, continuously differentiable function on $[0, T]$, satisfying:

$$\dot{x}(t) \le A(t)x(t) + B(t)$$

for each $t \in [0, T]$, where A, B are non-negative integrable functions on $[0, T]$, and vectors on the two side of the inequality are compared term by term. Show that:

$$x(t) \le \exp\left(\int_0^t A(\tau)\,d\tau\right) \cdot \left[x(0) + \int_0^t B(\tau)\,d\tau\right]$$

for each $t \in [0, T]$.

Solution:

We proceed by rewriting the inequality describing $\dot{x}(t)$ in integral form, and applying the Bellman-Gronwell Lemma:

$$\begin{aligned}
\dot{x}(t) &\leq A(t)x(t) + B(t) \\
\Rightarrow x(t) &\leq x(0) + \int_0^t \left[A(\tau)x(\tau) + B(\tau)\right] d\tau \\
&= \left[x(0) + \int_0^t B(\tau)\, d\tau\right] + \int_0^t A(\tau)x(\tau)\, d\tau \\
\Rightarrow x(t) &\leq \left[x(0) + \int_0^t B(\tau)\, d\tau\right] \cdot \exp\left(\int_0^t A(\tau)\, d\tau\right)
\end{aligned}$$

since the term in the square brackets is non-decreasing:

$$\frac{d}{dt}\left[x(0) + \int_0^t B(\tau)\, d\tau\right] = B(t) \geq 0$$

Example (**Discussion 4, Problem 4**). Suppose that the dynamical system:

$$\begin{aligned}
\dot{x}(t) &= A(t)x(t) + B(t)u(t) \\
y(t) &= C(t)x(t) + D(t)u(t) \\
x(t_0) &= x_0
\end{aligned}$$

admits the unique solution:

$$x(t) = \Phi(t, t_0)x_0 + \int_{t_0}^t \Phi(t, t_0)\, B(\tau)u(\tau)\, d\tau,$$

for each $t \in [t_0, \infty)$. Identify the state transition function, the output read-out map, and the response function.

Solution:

The state transition function $s(t, t_0, x_0, u[t_0, t])$ is:

$$x(t) = \Phi(t, t_0)x_0 + \int_{t_0}^t \Phi(t, t_0)\, B(\tau)u(\tau)\, d\tau$$

the output read-out map is:

$$y(t) = C(t)x(t) + D(t)u(t)$$

while the response function is:

$$y(t) = C(t)\Phi(t, t_0)x_0 + C(t) \cdot \int_{t_0}^t \Phi(t, t_0)\, B(\tau)u(\tau)\, d\tau + D(t)u(t)$$

3.3 Lecture 8

A system representation is a mathematical model of an input-output system. Consider, for example, the representation in Lecture 7:

$$\begin{aligned} \dot{x} &= f(x,u,t), \qquad f : \mathbb{R}^n \times \mathbb{R}^{n_i} \times \overline{\mathbb{R}^+} \to \mathbb{R}^n \\ y &= h(x,u,t), \qquad h : \mathbb{R}^n \times \mathbb{R}^{n_i} \times \overline{\mathbb{R}^+} \to \mathbb{R}^{n_0} \end{aligned}$$

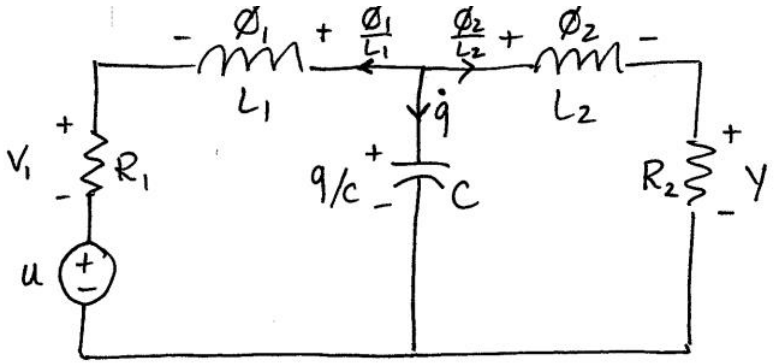
with initial condition $x(t_0) = x_0$.

Another example is the discrete-time representation:

$$\begin{aligned} x_{k+1} &= f(x_k, u_k, k) \\ y_k &= h(x_k, u_k, k) \end{aligned}$$

To motivate the following abstract description of a dynamical system, consider first the concrete example of a passive electrical circuit.

Example (**Electrical Circuit**). As an example, consider the following electrical circuit:



where:

1. ϕ_1, ϕ_2, q: Fluxes ϕ_1, ϕ_2 and capacitor charge q, considered here to be *state variables*.
2. u: Voltage source, considered here to be the *input*
3. y: Voltage across R_2, considered here to be the *output*

Then, by KCL and KVL, we have:

$$\begin{aligned} \dot{q} &= \frac{1}{L_1}\phi_1 - \frac{1}{L_2}\phi_2 \\ \dot{\phi_1} &= \frac{1}{C}q - \frac{R_1}{L_1}\phi_1 - u \\ \dot{\phi_2} &= \frac{1}{C}q - \frac{R_2}{L_2}\phi_2 \\ y &= \frac{R_1}{L_2}\phi_2 \end{aligned}$$

In system representation:

$$\begin{bmatrix} \dot{q} \\ \dot{\phi}_1 \\ \dot{\phi}_2 \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{L_1} & -\frac{1}{L_2} \\ \frac{1}{C} & -\frac{R_1}{L_1} & 0 \\ \frac{1}{C} & 0 & -\frac{R_2}{L_2} \end{bmatrix} \begin{bmatrix} q \\ \phi_1 \\ \phi_2 \end{bmatrix} + \begin{bmatrix} 0 \\ -1 \\ 0 \end{bmatrix} u$$

$$\Rightarrow y = \begin{bmatrix} 0 & 0 & \frac{R_2}{L_2} \end{bmatrix} \begin{bmatrix} q \\ \phi_1 \\ \phi_2 \end{bmatrix}$$

The above description can be generalized into a formal definition of a dynamical system

Definition 3.7 (**Dynamical System**). *Let **time** T be a variable defined on $\mathcal{T} = (\infty, \infty)$ or $[0, \infty)$ (continuous-time case) or $\{nT, n \in \mathbb{Z}\}$ (discrete-time case). A **dynamical system** is a 5-tuple:*

$$(\mathcal{U}, \Sigma, \mathcal{Y}, s, r)$$

defined on $\mathcal{T}$, where:

1. **Input space:** $(\mathcal{U})$

   $\mathcal{U}$ is the set of input functions *from $\mathcal{T} \to U$:*

   $$u(t) = \{u(t), \mathcal{T} \to \mathcal{U}\}$$

   Typically, $\mathcal{U} = \mathbb{R}^{n_i}$.

2. **Output space:** $(\mathcal{Y})$

   $\mathcal{Y}$ is the set of output functions *from $\mathcal{T} \to \mathcal{Y}$*

   $$u(t) = \{y(t), \mathcal{T} \to \mathcal{Y}\}$$

   Typically, $\mathcal{Y} = \mathbb{R}^{n_o}$.

3. **States:** (Σ)

   Σ is a set, called the state space*, that contains all* state trajectories

   $$\Sigma = \{x(t), t \in \mathcal{T}\}$$

   Typically, $\Sigma = \mathbb{R}^n$.

4. **State transition map:** (s)

   The state transition map*:*

   $$s : \mathcal{T} \times \mathcal{T} \times \Sigma \times \mathcal{U} \to \Sigma$$

   is a mapping from a given pair of initial and final times *(t_0, t_1), an* initial state *(x_0), and an* input functions *$(u[t_0, t_1])$, to a* final state *$(x(t_1))$:*

   $$x(t_1) = s(t_1, t_0, x_0, u)$$

By convention, the third element in $s(\cdot,\cdot,\cdot,\cdot)$ refers to the location of the trajectory $x(t)$ at the time given by the second element in $s(\cdot,\cdot,\cdot,\cdot)$. For instance, writing $x(t) = s(t,\tau,x',u)$ indicates that the initial condition under consideration is $x' = x(\tau)$.

5. **Output Read-out Map:** (r)

   The output readout function*:*

$$r : \mathcal{T} \times \Sigma \times \mathcal{U} \to \mathcal{Y}$$

   is a mapping from a given point in time t, and the state and output at time t, to the output $y(t)$ at time t:

$$y(t) = r(t, x(t), u(t))$$

Remark (**Difference between $u[t_0,t_1]$ and $u(t)$**). All state transition maps s take into account the action of u on the state trajectory $x(t)$ *at all times in* $\mathcal{T}$. When considering state transitions between different points in time (given an initial state), it is usually *insufficient to merely consider the effect of u at any specific point in time* between $[t_0,t_1]$. On the other hand, in our model of dynamical systems, we assume that *the output $y(t)$ at any point in time t' depends only on the state and input at time t'* (i.e. $x(t'), u(t')$, respectively. Thus, when calculating $y(t')$, it is enough to know $x(t')$ and $u(t')$; we do not need to know $u[t_0,t_1]$.

Definition 3.8. *For any given dynamical system, the state transition map s is required to satisfy the following two axioms:*

1. **State Transition Axiom:**

   Let $t_0, t_1 \in \mathcal{T}$ be given, with $t_0 \leq t_1$. The state transition axiom *states that if $u(t), \tilde{u}(t) \in \mathcal{U}$ satisfy:*

$$u(t) = \tilde{u}(t), \qquad \forall t \in [t_0,t_1] \cap \mathcal{T}$$

   for each interval $[t_0,t_1] \cap \mathcal{T}$, then:

$$s(t_1, t_0, x_0, u) = s(t_1, t_0, x_0, u)$$

2. **Semi-Group Axiom:**

   Let $t_0, t_1, t_2 \in \mathcal{T}$ be given, with $t_0 \leq t_1 \leq t_2$. The semi-group axiom *states that if $u(t), \tilde{u}(t) \in \mathcal{U}$ satisfy:*

$$\begin{aligned} s(t_2, t_1, x(t_1), u) &= s(t_2, t_1, s(t_1, t_0, x_0, u), u) \\ &= s(t_2, t_0, x_0, u) \end{aligned}$$

   for each initial state $x_0 \in \Sigma$ and input function $u \in \mathcal{U}$.

Remark (**Interpretations of the State Transition Axioms**).

1. State Transition Axiom:

   Only the input between the initial and final points of time will affect the state trajectory. This property is suggested by writing:

$$x(t_1) = s(t, t_0, x_0, u[t_0, t_1])$$

2. Semi-Group Axiom:

   In other words, the concatenated effects of applying the same input u throughout the time intervals $[t_0, t_1]$ and $[t_1, t_2]$ is the same as the effect of applying u throughout the total time interval $[t_0, t_2]$, assuming all initial conditions are aligned.

Below, we define two important classes of dynamical systems—*time-invariant* systems and *linear* dynamical systems.

Definition 3.9 (**Shift Operator**). *Define the* ***shift operator*** $T_\tau : \mathcal{U} \to \mathcal{U}$ *as:*

$$(T_\tau u)(t) = u(t - \tau)$$

(Similar notations are used for $T_\tau : \mathcal{Y} \to \mathcal{Y}$*.)*

Definition 3.10 (**Time-Invariant Dynamical System**). *A dynamical system is said to be* ***time-invariant*** *if:*

1. $\mathcal{U}$ *is closed under* T_τ*, for each* τ*.*
2. *For each* $t_0, t, \tau \in \mathcal{T}$*, where* $t_0 \leq t_1$*, and each* $x_0 \in \Sigma, u \in \mathcal{U}$*, we have:*

$$s(t_1, t_0, x_0, u) = s(t_1 + \tau, t_0 + \tau, x_0, T_\tau u)$$

Definition 3.11 (**Linear Dynamical Systems**). *A dynamical system is said to be* ***linear*** *if:*

1. $\mathcal{U}, \Sigma, \mathcal{Y}$ *are linear spaces over the same field* $\mathbb{F}$*.*
2. *For each* $t_0, t \in \mathcal{T}$*, with* $t_0 \leq t$*, the response map* ρ *is linear in* $\Sigma \times \mathcal{U}$*, i.e. for any* $x_1, x_2 \in \Sigma$ *and* $\alpha_1, \alpha_2 \in \mathbb{F}$*:*

$$\begin{aligned} &\rho(t, t_0, \alpha_1 x_1 + \alpha_2 x_2, \alpha_1 u_1 + \alpha_2 u_2) \\ =&\alpha_1 \cdot \rho(t, t_0, x_1, u_1) + \alpha_2 \cdot \rho(t, t_0, x_2, u_2) \end{aligned}$$

Remark. The above axioms in the definition of linearity for dynamical systems basically state that, in order for a dynamical system to be considered linear, the parameters and mappings associated with its definition must satisfy the following properties:

1. Any two states, or their linear combinations, can be added together. So can any two inputs or any two outputs.

2. For any given initial time t_0 and final (time of response) time t, the (output) response of the dynamical system to a linear combination of 2-tuples of (initial state, input), i.e.:

$$\alpha_1(x_{01}, u_1) + \cdots + \alpha_n(x_{0n}, u_n)$$

is the linear combination of the responses of the system to any particular 2-tuple (x_{n1}, u_n), with the same coefficients $\alpha_1, \cdots, \alpha_n$.

Definition 3.12 (**Zero-State Response, Zero-Input Response**). *Given a linear dynamical system $(\mathcal{U}, \Sigma, \mathcal{Y}, r, s)$, suppose θ_Σ and $\theta_\mathcal{U}$ are the zero elements of Σ and $\mathcal{U}$, respectively. Then:*

$$\begin{aligned} (x_0, u) &= (\theta_\Sigma, u) + (x_0, \theta_\mathcal{U}) \\ \Rightarrow \rho(t, t_0, x_0, u) &= \rho(t, t_0, \theta_\Sigma, u) + \rho(t, t_0, x_0, \theta_\mathcal{U}) \end{aligned}$$

Define:

$$\begin{aligned} \rho(t, t_0, \theta_\Sigma, u) &\equiv \textbf{zero-state response} \\ \rho(t, t_0, x_0, \theta_\mathcal{U}) &\equiv \textbf{zero-input response} \end{aligned}$$

We have thus shown that the response map of any linear dynamical system can be written as the sum of a zero-state response and a zero-input response. The linearity of each can be established by observing that both θ_Σ and $\theta_\mathcal{U}$ can be written as linear combinations of itself:

$$\begin{aligned} &\rho(t, t_0, \theta_\Sigma, \alpha_1 u_1 + \alpha_2 u_2) \\ =&\alpha_1 \cdot \rho(t, t_0, \theta_\Sigma, u_1) + \alpha_2 \cdot \rho(t, t_0, \theta_\Sigma, u_2) \\ &\rho(t, t_0, \alpha_1 x_{01} + \alpha_2 x_{02}, \theta_\mathcal{U}) \\ =&\alpha_1 \cdot \rho(t, t_0, x_{01}, \theta_\mathcal{U}) + \alpha_2 \cdot \rho(t, t_0, x_{02}, \theta_\mathcal{U}) \end{aligned}$$

Finally, we establish the concept of *equivalent states* and *equivalent representations* for two dynamical systems.

Definition 3.13 (**Equivalent States**). *Let $D = (\mathcal{U}, \Sigma, \mathcal{Y}, s, r)$ and $\tilde{D} = (\mathcal{U}, \tilde{\Sigma}, \mathcal{Y}, \tilde{s}, \tilde{r})$ be two dynamical systems with the same input and output spaces ($\mathcal{U}$ and $\mathcal{Y}$, respectively). We say that an initial state $x_0 \in \Sigma$ of D and an initial state $\tilde{x}_0 \in \tilde{\Sigma}$ of $\tilde{D}$ are **equivalent** if, for each $t \geq t_0$:*

$$\rho(t, t_0, x_0, u[t_0, t]) = \rho(t, t_0, \tilde{x}0, u[t_0, t])$$

In other words, stating that $x_0 \in \Sigma$ and $\tilde{x} \in \tilde{\Sigma}$ are equivalent states is the same as stating the following—If at time t_0, the systems D and $\tilde{D}$ are initialized at state x_0 and $\tilde{x}_0$, respectively, and are subject to the same input $u[t_0, t]$ during the interval $[t_0, t]$, then they must have the same response ρ at t.

Definition 3.14 (**Equivalent Representations**). *Two dynamical systems D and $\tilde{D}$ are said to be equivalent if and only if, for each $t_0 \in T, x \in D$, there exists at least one state $\tilde{x} \in \tilde{D}$ that is equivalent to x at t_0. Thus, equivalent system representaitons have the same input-output pairs.*

3.4 Lecture 8 Discussion

Example (**Discussion 4, Problem 4**). Suppose that the dynamical system:

$$\begin{aligned} \dot{x}(t) &= A(t)x(t) + B(t)u(t) \\ y(t) &= C(t)x(t) + D(t)u(t) \\ x(t_0) &= x_0 \end{aligned}$$

admits the unique solution:

$$x(t) = \Phi(t, t_0)x_0 + \int_{t_0}^{t} \Phi(t, t_0)\, B(\tau)u(\tau)\, d\tau,$$

for each $t \in [t_0, \infty)$. Identify the state transition function, the output read-out map, and the response function.

Solution:

The state transition function $s(t, t_0, x_0, u[t_0, t])$ is:

$$x(t) = \Phi(t, t_0)x_0 + \int_{t_0}^{t} \Phi(t, t_0)\, B(\tau)u(\tau)\, d\tau$$

the output read-out map is:

$$y(t) = C(t)x(t) + D(t)u(t)$$

while the response function is:

$$y(t) = C(t)\Phi(t, t_0)x_0 + C(t) \cdot \int_{t_0}^{t} \Phi(t, t_0)\, B(\tau)u(\tau)\, d\tau + D(t)\, u(t)$$

3.5 Lecture 9

In Sections 7 and 8, we considered the system to be evolving according the differential equation $\dot{x} = f(x,t)$, where f can be of any form, so long as it guarantees the existence of a unique solution. Here, we restrict our attention to specific forms for $f(x,t)$, as well as specific forms for writing the output in terms of the states and inputs. This will allow us to define a mapping $\Phi(t,t_0)$, called the *state transition matrix*, that concisely describes the state trajectory $x(t)$.

Definition 3.15 (**System Representation**). *The system representation* $R = [A(\cdot), B(\cdot), C(\cdot), D(\cdot)]$ *stands for the dynamical system:*

$$\dot{x}(t) = A(t)x(t) + B(t)u(t), \qquad x(t_0) = x_0 \tag{3.2}$$
$$y(t) = C(t)x(t) + D(t)u(t) \tag{3.3}$$

where we have:

$$u(t) \in \mathbb{R}^{n_i}$$
$$y(t) \in \mathbb{R}^{n_0}$$
$$x(t) \in \mathbb{R}^{n},$$

with matrix-valued piecewise continuous *functions:*

$$A(t) \in \mathbb{R}^{n\times n}$$
$$B(t) \in \mathbb{R}^{n\times n_i}$$
$$C(t) \in \mathbb{R}^{n_o\times n}$$
$$D(t) \in \mathbb{R}^{n_o\times n_i}$$

The input $u(t) \in \mathcal{U}$, where $\mathcal{U}$ is the set of piecewise continuous functions from $\overline{\mathbb{R}^+} \to \mathbb{R}^{n_i}$.

We wish to show that (3.2) and (3.3) satisfy the conditions of the existence and uniqueness theorem for differential equations; this would imply that $x(t)$ and $y(t) \in \mathbb{R}^{n_0}$ is well-defined for all $t \geq t_0$. Rewrite (3.2) as:

$$\begin{aligned}\dot{x}(t) &= A(t)x(t) + B(t)u(t)\\ &\equiv p(x(t),t)\end{aligned}$$

Observe that for each x, the function $p(x,t)$ is piecewise continuous in t, since $A(t), B(t), u(t)$ are piecewise continuous in t. Also, for each fixed t, $p(x,t)$ is *globally* Lipschitz in x (with Lipschitz constant $\|A(t)\|_i$ piecewise continuous in t), as shown below:

$$\begin{aligned}|p(x_1,t) - p(x_2,t)| &= |A(t)(x_1 - x_2)\\ &\leq \|A(t)\|_i |x_1 - x_2|\end{aligned}$$

The solution $x(t)$ can be represented in terms of the *state transition map* s, while the output can be represented by the *response map* ρ:

$$x(t) = s(t, t_0, x_0, u[t_0, t])$$
$$y(t) = \rho(t, t_0, x_0, u[t_0, t])$$

Example (**Linearization**). One of the reasons for adopting a system representation of the form (3.2), (3.3), is that state perturbations in non-linear systems can be characterized via linearization. Consider a general non-linear system with dynamics of the form:

$$\dot{x} = f(x, u, t), \qquad x(t_0) = x_0$$
$$y = h(x, u, t)$$

and suppose that a small perturbation is applied to the state (x) and input (u) of the system, thus inducing a slight change in the output (y):

$$x \longrightarrow x + \delta x$$
$$x(t_0) \longrightarrow x_0 + \delta x_0$$
$$u \longrightarrow u + \delta u$$
$$y \longrightarrow y + \delta y$$

Then, Taylor expansion gives us:

$$\begin{aligned}
\dot{x} + \delta\dot{x} &= f(x + \delta x, u + \delta u, t) \\
&= f(x, u, t) + \underbrace{\frac{\partial}{\partial x} f(x, u, t)\bigg|_{x,u}}_{\equiv A(t) \in \mathbb{R}^{n \times n}} \delta x + \underbrace{\frac{\partial}{\partial u} f(x, u, t)\bigg|_{x,u}}_{\equiv B(t) \in \mathbb{R}^{n \times n_i}} \delta u \\
\Rightarrow \delta\dot{x} &= A(t) \cdot \delta x + B(t) \cdot \delta u
\end{aligned}$$

Similarly, we have:

$$\begin{aligned}
y + \delta y &= h(x + \delta x, u + \delta u, t) \\
&= h(x, u, t) + \underbrace{\frac{\partial}{\partial x} h(x, u, t)\bigg|_{x,u}}_{\equiv C(t) \in \mathbb{R}^{n_o \times n}} \delta x + \underbrace{\frac{\partial}{\partial u} f(x, u, t)\bigg|_{x,u}}_{\equiv D(t) \in \mathbb{R}^{n_o \times n_i}} \delta u \\
\Rightarrow \delta y &= C(t) \cdot \delta x + D(t) \cdot \delta u
\end{aligned}$$

In summary, we have:

$$\delta\dot{x} = A(t) \cdot \delta x + B(t) \cdot \delta u$$
$$\delta y = C(t) \cdot \delta x + D(t) \cdot \delta u$$

Below, we define the *state transition matrix* for dynamical systems with system representation of the form (3.2) and (3.3), and demonstrate its relationship to the state $x(t)$.

Definition 3.16 (**State Transition Matrix**). *Consider a dynamical system $(\mathcal{U}, \Sigma, \mathcal{Y}, s, r)$ with system representation given as in* (3.2), (3.3). *The* ***state transition matrix*** *$\Phi(t, t_0)$ of the system is defined as the unique solution to the following differential equation, where $X(t) \in \mathbb{R}^{n\times n}$ for each $t \in \mathbb{R}$:*

$$\dot{X} = A(t)X, \qquad X(t_0) = X_0 \tag{3.4}$$

Proposition 3.17. *Consider a dynamical system $(\mathcal{U}, \Sigma, \mathcal{Y}, s, r)$ with system representation given as in* (3.2), (3.3), *and state transition matrix $\Phi(t, t_0)$. Then:*

1. *The (unique) solution to the differential equation:*

$$\dot{x} = A(t)x, \qquad x(t_0) = x_0$$

is given by:

$$x(t) \equiv s(t, t_0, x_0) = \Phi(t, t_0)x_0$$

Moreover, for each $t, t_0, t_1 \in \overline{\mathbb{R}^+}$, we have:

2. $\Phi(t, t_0) = \Phi(t, t_1) \cdot \Phi(t_1, t_0)$.
3. $[\Phi(t, t_0)]^{-1} = \Phi(t_0, t)$
4. $det(t, t_0) = \exp\left(\int_{t_0}^{t} tr(A(\tau))\, d\tau\right)$

Proof.

1. By the Fundamental Theorem, we only have to show that the given expression satisfies the given differential equation and initial conditions. For the differential equation, we have:

$$\frac{d}{dt}\big[\Phi(t, t_0)x_0\big] = \left[\frac{d}{dt}\Phi(t, t_0)\right] x_0 = A(t)x_0$$

At $t = t_0$, we have:

$$x(t_0) = \Phi(t_0, t_0)x_0 = x_0$$

The proof is done.

2. Here, we will show that the expression on the right-hand side satisfies the matrix differential equation that defines $\Phi(t, t_0)$, as well as the given initial condition:

$$\frac{d}{dt}\big[\Phi(t, t_1) \cdot \Phi(t_1, t_0)\big] = \big[A(t) \cdot \Phi(t, t_1)\big] \cdot \Phi(t_1, t_0),$$

since, by definition, $\Phi(t, t_1)$ is the unique solution to the matrix equation:

$$\dot{X} = A(t)\,X, \qquad X(t_1) = I$$

For the initial condition, we have at t_1:

$$\Phi(t_1, t_1)\Phi(t_1, t_0) = \Phi(t_1, t_0)$$

3. Observe that:

$$I = \Phi(t_0, t_0) = \Phi(t_0, t) \cdot \Phi(t, t_0)$$
$$I = \Phi(t, t) = \Phi(t, t_0) \cdot \Phi(t_0, t),$$

so $\Phi(t_0, t)^{-1} = \Phi(t, t_0)$.

4. The given equation can be rewritten in its differential form, as follows:

$$\frac{d}{dt}\Phi(t, t_0) = tr(A) \cdot \det\big(\Phi(t, t_0)\big)$$

This can be shown by observing that:

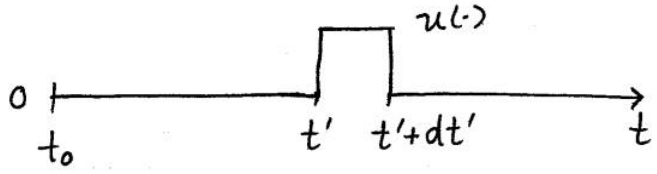
$$\begin{aligned}
\Phi(t, t_0) &= \Phi(t, t_0) + A(t)\,\Phi(t_0, t)\,dt + O(dt^2) \\
&= (I + A(t)dt)\Phi(t, t_0) + O(dt^2) \\
\Rightarrow \det\big(\Phi(t + dt, t_0)\big) &= \left[1 + \sum_{i=1}^{n} a_{ii}dt + O(dt^2)\right] \cdot \det\big(\Phi(t, t_0)\big) + O(dt^2) \\
&= [1 + tr(A) \cdot dt] \cdot \det\Phi(t, t_0) + O(dt^2) \\
\Rightarrow \frac{d}{dt}\det\Phi(t, t_0) &= \lim_{t \to 0} \frac{\det\Phi(t + dt, t_0) - \det\Phi(t, t_0)}{dt} = tr(A) \cdot \det(\Phi(t, t_0))
\end{aligned}$$

■

Now, we wish to demonstrate what happens when a nonzero input function $u(t)$ is superimposed onto the system. Consider an input $u(t)$ imposed onto the system during the infinitesimal time period $[t', t' + dt]$, where:

$$t_0 \ll t' < t' + dt' \ll t$$

as shown in the figure below.



Analyzing the trajectory $x(t)$ at times $t', t' + dt$ and t, we find:

$$\begin{aligned}
x(t') &= \Phi(t', t_0)x_0 \\
\Rightarrow x(t' + dt') &= x(t') + \left[A(t')x(t') + B(t')u(t')\right] dt' \\
\Rightarrow x(t) &= \Phi(t, t' + dt')\, x(t' + dt') \\
&= \Phi(t, t' + dt') \left[x(t') + \left[A(t')x(t') + B(t')u(t')\right] dt'\right] \\
&= \Phi(t, t' + dt') \left[I + A(t')\, dt'\right] x(t') + \Phi(t, t' + dt')\, B(t')u(t')\, dt' \\
&\approx \Phi(t, t' + dt')\, \Phi(t' + dt')
\end{aligned}$$

where $\frac{d}{dt}\Phi(t, t_0) = A(t)\,\Phi(t, t_0)$:

$$\begin{aligned}
A(t)\Phi(t, t_0) &= \frac{d}{dt}\Phi(t, t_0) \\
&\approx \frac{\Phi(t' + dt', t') - \Phi(t', t')}{dt'} \\
&= \frac{\Phi(t' + dt', t') - I}{dt'} \\
\Rightarrow \Phi(t' + dt', t') &\approx I + A(t)\Phi(t, t_0)\, dt'
\end{aligned}$$

Theorem 3.18. *The state transition and response maps for a dynamical system with system representation* (3.2), (3.3) *are:*

$$s(t, t_0, x_0, u[t_0, t]) = \Phi(t, t_0)x_0 + \int_{t_0}^{t} \Phi(t, t')B(t')u(t')\, dt', \tag{3.5}$$

$$\rho(t, t_0, x_0, u[t_0, t]) = C(t)\Phi(t, t_0)x_0 + C(t) \int_{t_0}^{t} \Phi(t, t')B(t')u(t')\, dt' + D(t)u(t), \tag{3.6}$$

respectively.

Proof. To verify (3.5), we must demonstrate that the expression on its right-hand side satisfies the differential equation (3.2). By taking the derivative of s with respect to time, we have:

$$\begin{aligned}
\frac{d}{dt}s(t, t_0, x_0, u[t_0, t]) &= \left[\frac{d}{dt}\Phi(t, t_0)\right] x_0 + \Phi(t, t)B(t)u(t) \\
&= \left[A(t)\Phi(t, t_0)\right] x_0 + B(t)u(t) \\
&= A(t)x(t) + B(t)u(t)
\end{aligned}$$

It remains to show that the initial condition is satisfied:

$$x(t_0) = \Phi(t, t_0)x_0 + 0 = x_0$$

Finally, (3.6) follows when (3.5) is substituted as $x(t)$ into (3.3). This completes the proof. ■

Theorem 3.19. *The state transition function* (3.5) *satisfies the state transition axiom and the semi-group axiom.*

Proof.

1. State Transition Axiom:

   Let inputs $u(\cdot)$ and $\overline{u}(\cdot)$ be given such that they take identical values in the time interval $\tau \in [t_0, t]$. Then we have:

$$\begin{aligned} s(t, t_0, x_0, u[t_0, t]) &= \Phi(t, t_0)x_0 + \int_{t_0}^{t} \Phi(t, t')B(t')u(t')\,dt' \\ &= \Phi(t, t_0)x_0 + \int_{t_0}^{t} \Phi(t, t')B(t')\overline{u}(t')\,dt' \\ &= s(t, t_0, x_0, \overline{u}[t_0, t]) \end{aligned}$$

2. Semi-Group Axiom:

   To check the semi-group axiom, we must compare $s(t_2, t_1, s(t_1, t_0, x_0, u[t_0, t_1], u[t_1, t_2]))$ and $s(t_2, t_0, x_0, u[t_0, t_2])$:

$$\begin{aligned} & s(t_2, t_1, s(t_1, t_0, x_0, u[t_0, t_1], u[t_1, t_2])) \\ =& \Phi(t_2, t_1) \cdot \left(\Phi(t_1, t_0)x_0 + \int_{t_0}^{t_1} \Phi(t_1, \tau)B(\tau)u(\tau)d\tau \right) + \int_{t_1}^{t_2} \Phi(t_2, \tau)B(\tau)u(\tau)d\tau \\ =& \Phi(t_2, t_0)x_0 + \int_{t_0}^{t_1} \Phi(t_2, \tau)B(\tau)u(\tau)d\tau + \int_{t_1}^{t_2} \Phi(t_2, \tau)B(\tau)u(\tau)d\tau \\ =& \Phi(t_2, t_0)x_0 + \int_{t_0}^{t_2} \Phi(t_2, \tau)B(\tau)u(\tau)d\tau \end{aligned}$$

■

3.6 Lecture 9 Discussion

Example (**Discussion 5, Problem 5, Fall 2009, 2015 Midterms**). For a non-singular $M(t) \in \mathbb{R}^{n\times n}$, determine an expression for:

$$\frac{d}{dt}\left[M^{-1}(t)\right]$$

in terms of $\dot{M}(t)$ and $M^{-1}(t)$.

Solution :

The desired result can be derived by applying the definition for the inverse of a matrix and the product rule for differentiation:

$$\begin{aligned} & I = M(t)\cdot M^{-1}(t) \\ \Rightarrow & O = \left[\frac{d}{dt}M(t)\right] M^{-1}(t) + M(t)\left[\frac{d}{dt}M^{-1}(t)\right] \\ \Rightarrow & \frac{d}{dt}M^{-1}(t) = -M^{-1}(t)\left[\frac{d}{dt}M(t)\right] M^{-1}(t) \end{aligned}$$

Example (**Discussion 5, Problems 6, 7, Fall 2014 Midterm**). Given a system of the form (3.2) and (3.3), with state transition matrix $\Phi(t, t_0)$

1. Find an expression for:
$$\frac{d}{d\tau}\Phi(t,\tau)$$
in terms of $\Phi(t, t_0)$ and $A(t)$.

2. Prove that $\Phi(t_0, t)$ is the unique solution to the matrix differential equation:
$$\frac{d}{dt}X(t) = -X(t)\,A(t), \qquad A(t_0) = I$$

Solution:

1. We repeat the solution process of the above example for $\Phi(t,\tau)$:

$$\begin{aligned} & I = \Phi(t,t) = \Phi(t,\tau)\,\Phi(\tau,t) \\ \Rightarrow & O = \left[\frac{d}{d\tau}\Phi(t,\tau)\right]\Phi(t,\tau) + \Phi(t,\tau)\left[\frac{d}{d\tau}\Phi(\tau,t)\right] \\ \Rightarrow & \frac{d}{d\tau}\Phi(t.\tau) = -\Phi(t,\tau)\left[\frac{d}{dt}\Phi(\tau,t)\right]\Phi(t,\tau) \\ & = -\Phi(t,\tau)\left[A(\tau)\,\Phi(\tau,t)\right] \\ & = -\Phi(t,\tau)A(\tau) \end{aligned}$$

where $\Phi(\tau, t)$ is the unique solution:

$$\frac{d}{d\tau}X = A(\tau)X, \qquad X(\tau) = I.$$

2. First, we verify that $\Phi(t_0, t)$ satisfies the given differential equation. This follows directly from the results derived above (simply replace t and τ with t_0 and t, respectively:

$$\begin{aligned} &\because \frac{d}{d\tau}\Phi(t.\tau) = -\Phi(t,\tau)A(\tau) \\ &\Rightarrow \frac{d}{dt}\Phi(t_0,t) = -\Phi(t_0,t)\,A(t) \end{aligned}$$

Remark. Any uniqueness issues in the definition of $\Phi(t, t_0)$ can be circumvented by noting that, when $A(t)$ is piecewise continuous in t, the differential equation $\frac{d}{dt}X(t) = A(t)X(t)$ satisfies the Fundamental Theorem (which extends to matrix differential equations).

Example (**Discussion 5, Problem 8**). Calculate the state transition matrix for the differential equation $\dot{x}(t) = A(t)x(t)$, where:

$$A(t) = \begin{bmatrix} t & 2 \\ 0 & -1 \end{bmatrix}$$

Solution-1:

Since $A(t) \in \mathbb{R}^{2\times 2}$ for each $t \in \mathbb{R}$, the state space Σ is 2-dimensional, i.e. $x(t) = (x_1(t), x_2(t))^T$. Let the initial conditions be $x_1(t_0) = a$ and $x_2(t_0) = b$, and rewrite the differential equation as:

$$\begin{aligned} \dot{x}_1 &= tx_1 + 2x_2 \\ \dot{x}_2 &= -x_2 \end{aligned}$$

The second equation implies that $x_2(t) = be^{-(t-t_0)}$. Substituting this result into the first equation, we have:

$$\dot{x}_1 = tx_1 + 2be^{-(t-t_0)}$$

The solution to $x_1(t)$ can then be found via the integrating factor method:

$$\begin{aligned} &\dot{x}_1 = tx_1 + 2be^{-t} \\ \Rightarrow &e^{-\frac{1}{2}t^2}\dot{x}_1 - te^{-\frac{1}{2}t^2}x_1 = 2be^{-t-\frac{1}{2}t^2} \\ \Rightarrow &\frac{d}{dt}\left(e^{-\frac{1}{2}t^2}x_1\right) = 2be^{-t-\frac{1}{2}t^2} \\ \Rightarrow &e^{-\frac{1}{2}t^2}x_1 - e^{-\frac{1}{2}t_0^2}\cdot a = 2\int_{t_0}^{t} e^{-\tau-\frac{1}{2}\tau^2}d\tau \cdot b, \end{aligned}$$

where, in the final step, we have integrated from t_0 to t on both sides of the equality. In summary, we have for $x_1(t)$ and $x_2(t)$:

$$x_1(t) = e^{\frac{1}{2}(t^2-t_0^2)} \cdot a + 2e^{-\frac{1}{2}t_0^2} \int_{t_0}^{t} e^{-\tau-\frac{1}{2}\tau^2} d\tau \cdot b$$
$$x_2(t) = e^{-(t-t_0)} \cdot b$$

Since the state transition matrix uniquely satisfies $x(t) = \Phi(t, t_0)x_0$, we find that:

$$\Phi(t, t_0) = \begin{bmatrix} e^{\frac{1}{2}(t^2-t_0^2)} & 2e^{-\frac{1}{2}(t^2-t_0)} \int_{t_0}^{t} e^{-\tau-\frac{1}{2}\tau^2} d\tau \\ 0 & e^{-(t-t_0)} \end{bmatrix}$$

Solution-2:

Another method is to solve for $\Phi(t, t_0)$ directly, from its definition as the unique solution to the differential equation:

$$\dot{X}(t) = A(t)X(t), \qquad X(t_0) = I$$

Define the components of $\Phi(t, t_0)$ as:

$$\Phi(t, t_0) \equiv \begin{bmatrix} \Phi_{11}(t, t_0) & \Phi_{12}(t, t_0) \\ \Phi_{21}(t, t_0) & \Phi_{22}(t, t_0) \end{bmatrix}$$

Then the differential equation uniquely satisfied by $\Phi(t, t_0)$ becomes:

$$\begin{aligned} \begin{bmatrix} \dot{\Phi}_{11}(t, t_0) & \dot{\Phi}_{12}(t, t_0) \\ \dot{\Phi}_{21}(t, t_0) & \dot{\Phi}_{22}(t, t_0) \end{bmatrix} &= \begin{bmatrix} A_{11}(t) & A_{12}(t) \\ A_{21}(t) & A_{22}(t) \end{bmatrix} \begin{bmatrix} \Phi_{11}(t, t_0) & \Phi_{12}(t, t_0) \\ \Phi_{21}(t, t_0) & \Phi_{22}(t, t_0) \end{bmatrix} \\ &= \begin{bmatrix} t & 2 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} \Phi_{11}(t, t_0) & \Phi_{12}(t, t_0) \\ \Phi_{21}(t, t_0) & \Phi_{22}(t, t_0) \end{bmatrix} \end{aligned}$$

This can be expressed as four differential equations, one each for $\dot{\Phi}_{11}(t, t_0)$, $\dot{\Phi}_{12}(t, t_0)$, $\dot{\Phi}_{21}(t, t_0)$, and $\dot{\Phi}_{22}(t, t_0)$:

$$\dot{\Phi}_{11} = t\Phi_{11} + 2\Phi_{21}$$
$$\dot{\Phi}_{11} = t\Phi_{11} + 2\Phi_{22}$$
$$\dot{\Phi}_{21} = -\Phi_{21}$$
$$\dot{\Phi}_{22} = -\Phi_{22}$$

with initial conditions:

$$\dot{\Phi}_{11}(t_0, t_0) = 1$$
$$\dot{\Phi}_{12}(t_0, t_0) = 0$$
$$\dot{\Phi}_{21}(t_0, t_0) = 0$$
$$\dot{\Phi}_{22}(t_0, t_0) = 1$$

From the differential equations for Φ_{21} and Φ_{22}, we immediately conclude that $\Phi_{21}(t, t_0) = 0$ and $\Phi_{22}(t, t_0) = e^{-(t-t_0)}$. Substituting into the differential equations for Φ_{11} and Φ_{12}, we have:

$$\dot{\Phi}_{11} = t\Phi_{11} + 2e^{-(t-t_0)}$$
$$\dot{\Phi}_{12} = t\Phi_{12}$$

Using the initial conditions $\dot{\Phi}_{11}(t_0, t_0) = 1$ and $\dot{\Phi}_{12}(t_0, t_0) = 0$, we have:

$$\Phi_{11}(t, t_0) = e^{\frac{1}{2}(t^2 - t_0^2)}$$
$$\Phi_{12}(t, t_0) = 2e^{-\frac{1}{2}(t^2 - t_0)} \int_{t_0}^{t} e^{-\tau - \frac{1}{2}\tau^2} d\tau,$$

in agreement with the solution presented above.

Example (**Discussion 5, Problem 10**). Determine whether the dynamical system described by the following equations is linear:

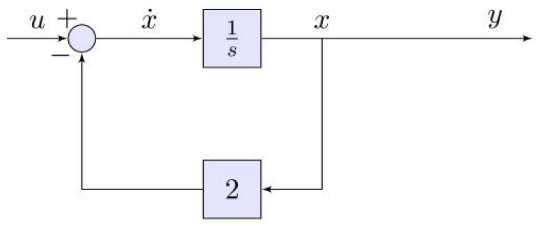
$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} \sin(x_2) \\ \sin(t)\, x_2(0) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(t)$$
$$y(t) = \begin{bmatrix} 0 & 0 \end{bmatrix} x(t)$$

Assume that $u(t)$ is piecewise continuous in t.

Solution:

The system is linear since the response map, which is identically 0, is linear with respect to any tuple of initial state and control, $(x_0, u[t_0, t])$. This problem emphasizes the fact that, even if the dynamics of the system appear non-linear, the system itself is linear by definition if its response map is linear.

Example (**Discussion 5, Problem 11**). Find a mathematical representation for the system presented below, and determine whether it is a dynamical system. Is it linear?



Solution:

1. Since multiplying by $1/s$ in the frequency domain is equivalent to an integrator in the time domain, we have:

$$\dot{x} = -2x + u$$
$$y = x$$

2. Applying (3.9) to the given differential equation, with the initial condition $x(t_0) = x_0$, we find:

$$x(t) = e^{-2(t-t_0)}x_0 + \int_{t_0}^{t} u(\tau) \cdot e^{-2(t-\tau)} d\tau$$

To verify that the above trajectory is that of a dynamical system, we must verify that it satisfies the state transition axiom and the semi-group axiom:

- State Transition Axiom:

  Let u and u' be input functions taking identical values at each point in $[t_0, t]$. Since the given state trajectory only depends on values of u in the time interval $[t_0, t]$, it satisfies state transition axiom.

- Semi-Group Axiom:

  We have:

$$\begin{aligned}
&s\big(t_2, t_1, s(t_1, t_0, x_0, u[t_0, t_1]), u[t_1, t_2]\big) \\
=&e^{-2(t_2-t_1)} \left(e^{-2(t-t_0)}x_0 + \int_{t_0}^{t_1} u(\tau) \cdot e^{-2(t_1-\tau)} d\tau \right) + \int_{t_1}^{t_2} u(\tau) \cdot e^{-2(t_2-\tau)} d\tau \\
=&e^{-2(t_2-t_0)}x_0 + \int_{t_0}^{t_1} u(\tau) \cdot e^{-2(t_1-\tau)} d\tau + \int_{t_1}^{t_2} u(\tau) \cdot e^{-2(t_1-\tau)} d\tau \\
=&e^{-2(t_2-t_0)}x_0 + \int_{t_0}^{t_2} u(\tau) \cdot e^{-2(t_1-\tau)} d\tau \\
=&s(t_1, t_0, x_0, u[t_0, t_2])
\end{aligned}$$

3.7 Lecture 10

It will become apparent that the solution to the differential equations (3.2), (3.2) involves the matrix exponential e^{tA}. Thus, we devote the first half of this section towards understanding this expression.

The following material, which establishes the convergence of e^{tA} for each square matrix A and each $t \in \mathbb{R}$, is not included in the Lecture 10 Notes. It originates from Professor Chee-Fai Yung's *Lecture Notes on Mathematical Control Theory* [12].

Below, we extend (3.5) and (3.6) to the special case where $A(t), B(t), C(t), D(t)$ are fixed matrices whose values are independent of time. However, we first require a result for the convergence of a particular sequence of matrix polynomials.

Definition 3.20 (Convergence of a Series of Matrices). *Let $\{c^{(k)}\}$ be a sequence of $m \times n$ matrices. We say that $\sum_k c_k$ converges if each series $\sum_k c_{ij}^{(k)}$ converges, where:*

$$c_k = [c_{ij}^{(k)}]_{m \times n}$$

If $\sum_k c_{ij}^{(k)}$ converges for each i, j, then we define:

$$[b_{ij}] \equiv \sum_k c_k$$

Theorem 3.21. *If $\sum_k \|c_k\|$ converges, then $\sum_k c_k$ converges.*

Proof. Without loss of generality, take the Frobenius norm. Since $|c_{ij}^{(k)}| \leq \|c_k\|$ for each i, j, and $\sum_k \|c_k\|$ converges, the series $\sum_k |c_{ij}^{\{}(k)\}$ converges for each i, j. In other words, $\sum_k c_{ij}^{(k)}$ converges (absolutely), for each i, j, so $\sum_k c_k$ converges by Theorem 3.20. ■

Theorem 3.22. *For each square matrix $\mathbf{A}$, the infinite sum:*

$$\sum_{k=0}^{\infty} \frac{\mathbf{A}^k}{k!}$$

converges.

Proof. For each $k = 0, 1, 2, \cdots$:

$$\left\| \frac{\mathbf{A}^k}{k!} \right\| = \frac{\|\mathbf{A}^k\|}{k!} \leq \frac{\|\mathbf{A}\|^k}{k!}$$

$$\Rightarrow \sum_{k=0}^{\infty} \left\| \frac{\mathbf{A}^k}{k!} \right\| \leq \sum_{k=0}^{\infty} \frac{\|\mathbf{A}\|^k}{k!} = e^{\|\mathbf{A}\|}$$

The right-hand side converges, by the Comparison Test. Thus, by Theorem 3.21, so does $\sum_{k=0}^{\infty} \frac{\boldsymbol{A}^k}{k!}$. ■

With the above results, we can now simplify (3.5) and (3.6) for the case where $A(t), B(t), C(t), D(t)$ are all constant.

Lecture 10 Notes begin here

Corollary 3.23. *Let D be a dynamical system with system representation (3.2) and (3.3), with $A(t), B(t), C(t), D(t)$ all constant (i.e. time-independent). Then the state transition matrix $\Phi(t, t_0)$ for D is:*

$$\Phi(t, t_0) = e^{(t-t_0)A} \tag{3.7}$$

As a result, the state transition and response maps for a dynamical system with system representation (3.2), (3.3), when $A(t), B(t), C(t), D(t)$ are all constant, are:

$$x(t) = e^{(t-t_0)A}x_0 + \int_{t_0}^{t} e^{(t-\tau)A}B\,u(\tau)\,d\tau \tag{3.8}$$

$$y(t) = C(t)\,e^{(t-t_0)A}x_0 + C(t)\int_{t_0}^{t} e^{(t-\tau)A}B\,u(\tau)\,d\tau + D\,u(t) \tag{3.9}$$

Proof. To show (3.7), we must show that $e^{(t-t_0)A}$ satisfies (3.4):

$$\begin{aligned}\frac{d}{dt}e^{(t-t_0)A} &= \frac{d}{dt}\left[\sum_{k=0}^{\infty}\frac{t^kA^k}{k!}\right] = \frac{d}{dt}\left[\sum_{k=0}^{\infty}\frac{t^kA^k}{k!}\right] = \sum_{k=0}^{\infty}\left[\frac{d}{dt}\left(\frac{t^kA^k}{k!}\right)\right] \\ &= \sum_{k=0}^{\infty}\frac{t^{k-1}A^k}{(k-1)!} = \sum_{k=1}^{\infty}\frac{t^kA^{k+1}}{k!} \\ &= Ae^{tA} = e^{tA}A\end{aligned}$$

The exchange of the infinite sum and the differentiation in the third equality relies on the fact that the convergence of $\left(\frac{k!}{t^kA^k}, k \geq 0\right)$ to e^{tA} is uniformly continuous in t (see Rudin [8]). (3.5) and (3.9) then follow by substituting (3.7) into (3.5) and (3.6). (Observe that the norm of A, a constant matrix, offers a natural choice for the Lipschitz constant. Thus, if $u(t)$ is piecewise continuous, all the conditions of the Fundamental Theorem are then satisfied.) ■

Other properties regarding the matrix exponential e^{tA} are discussed below.

Theorem 3.24 (Properties of the Matrix Exponential).

1. $e^{O} = I$.
2. $e^{(t+s)A} = e^{tA} \cdot e^{sA} = e^{sA} \cdot e^{tA}$.
3. *If (and only when) $AB = BA$, we have* $e^{t(A+B)} = e^{tA} \cdot e^{tB}$
4. $(e^{tA})^{-1} = e^{-tA}$
5. *If $A = PBP^{-1}$, then* $e^{tA} = Pe^{tB}P^{-1}$.

Proof.

1. By definition of the matrix exponential:

$$e^O = I + I + O + O + \cdots = I$$

2. By expanding the left-hand side of the given expression, we have:

$$\begin{aligned} e^{(t+s)A} &= \sum_{n=0}^{\infty} \frac{(t+s)^n A^n}{n!} = \sum_{n=0}^{\infty} \sum_{k=0}^{n} \binom{n}{k} t^k s^{n-k} \cdot \frac{A^n}{n!} \\ &= \sum_{k=0}^{\infty} \sum_{n=k}^{\infty} \frac{n!}{k!(n-k)!} t^k s^{n-k} \cdot \frac{A^n}{n!} \\ &= \sum_{k=0}^{\infty} \sum_{n=0}^{\infty} \frac{1}{k!n!} t^k s^n \cdot A^{(n+k)} \\ &= e^{tA} \cdot e^{sA} \\ &= e^{sA} \cdot e^{tA} \end{aligned}$$

3. " $\Rightarrow$ " By expanding the left-hand side of the given expression, we have:

$$\begin{aligned} e^{t(A+B)} &= \sum_{n=0}^{\infty} \frac{t^n (A+B)^n}{n!} = \sum_{n=0}^{\infty} \sum_{k=0}^{n} \binom{n}{k} A^k B^{n-k} \cdot \frac{t^n}{n!} \\ &= \sum_{k=0}^{\infty} \sum_{n=k}^{\infty} \frac{n!}{k!(n-k)!} A^k B^{n-k} \cdot \frac{t^n}{n!} \\ &= \sum_{k=0}^{\infty} \sum_{n=0}^{\infty} \frac{1}{k!n!} A^k B^n \cdot t^{(n+k)} \\ &= e^{tA} \cdot e^{tB} \\ &= e^{tB} \cdot e^{tA} \end{aligned}$$

where the fact that $AB = BA$ has been used, in conjunction with the binomial theorem, to separate the A and B terms in the expression $(A+B)^n$.

" $\Leftarrow$ " Conversely, suppose $e^{t(A+B)} = e^{tA}e^{tB}$. Differentiating both sides twice and taking $t = 0$, we have:

$$\begin{aligned} & e^{t(A+B)} = e^{tA} \cdot e^{tB} \\ \Rightarrow & (A+B)\, e^{t(A+B)} = e^{tA}(A+B)e^{tB} \\ \Rightarrow & (A+B)^2\, e^{t(A+B)} = e^{tA}\left(A(A+B) + (A+B)B\right) e^{tB} \\ \Rightarrow & A^2 + AB + BA + B^2 = A^2 + 2AB + B^2 \\ \Rightarrow & AB = BA \end{aligned}$$

4. Since A and $-A$ clearly commute, we have from the above derivation:

$$I = e^O = e^{t(A+(-A))} = e^{tA} \cdot e^{-tA} = e^{-tA} \cdot e^{tA}$$

5. By definition of the matrix exponential:

$$Pe^{tB}P^{-1} = P\left(\sum_{k=0}^{\infty} \frac{t^k B^k}{k!}\right) P^{-1} = \sum_{k=0}^{\infty} \frac{t^k P B^k P^{-1}}{k!}$$
$$= \sum_{k=0}^{\infty} \frac{t^k (PBP^{-1})^k}{k!} = \sum_{k=0}^{\infty} \frac{t^k A^k}{k!} = e^{tA}$$

■

Methods for calculating e^{tA}, when A is independent of t, include:

1. Direct Expansion
2. Laplace Transform
3. Cayley-Hamilton Theorem
4. Diagonalization or Jordan Canonical Form

However, if $A = A(t)$ is time-dependent, the differential equation must be directly solved (either for $x(t)$, or for $\Phi(t)$). The process is rather time-consuming. Examples are furnished in the discussion problems following this lecture.

For very simple matrices, e^{tA} can be directly calculated. In general, the results of the above theorem allow e^{tA} to be calculated via diagonalization or Jordan decomposition. Other methods for calculating e^{tA}, as listed above, are demonstrated in examples in Discussion Notes following this chapter.

Example. Suppose:

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

Find e^{tA} using the definition of the matrix polynomial.

Solution 1 : (Direct Calculation)

Since $A^2 = O$, we have $A^n = O$ for any $n = 2, 3, \cdots$. Thus:

$$e^{tA} = I + tA = \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix}$$

For more difficult examples, the Laplace transform can be used to find the matrix polynomial. Consider again the matrix differential equation $\dot{X} = AX, X(0) = I_n$. By taking the Laplace transform on both sides, we have:

$$\begin{aligned}
&\dot{X} = AX \\
\Rightarrow &s\hat{X}(s) - X(0) = A\hat{X}(s) \\
\Rightarrow &(sI - A)\hat{X}(s) = I \\
\Rightarrow &\hat{X}(s) = (sI - A)^{-1} \\
\Rightarrow &x(t) = \mathcal{L}^{-1}\{(sI - A)^{-1}\}
\end{aligned}$$

where the *adjugate matrix* $\text{adj}(sI - A)$ *of* A is the transpose of the *cofactor matrix* C *of* A, as shown below. We define A_{ij} as the matrix obtained by deleting the i-th row and j-th column of A:

$$\begin{aligned}
C_{ij}(A) &= (-1)^{i+j} det(A_{ij}) \\
adj(A) &= \big[C(A)\big]^T
\end{aligned}$$

Solution 2 : (Laplace Transform)

From above, we know that $e^{tA} = \mathcal{L}^{-1}(sI - A)$, so:

$$\begin{aligned}
&sI - A = \begin{bmatrix} s & -1 \\ 0 & s \end{bmatrix} \\
\Rightarrow &(sI - A)^{-1} = \frac{1}{s^2}\begin{bmatrix} s & 1 \\ 0 & s \end{bmatrix} = \begin{bmatrix} \frac{1}{s} & \frac{1}{s^2} \\ 0 & \frac{1}{s} \end{bmatrix} \\
\Rightarrow &e^{tA} = \mathcal{L}\{(sI - A)^{-1}\} = \mathcal{L}^{-1}\left\{\begin{bmatrix} \frac{1}{s} & \frac{1}{s^2} \\ 0 & \frac{1}{s} \end{bmatrix}\right\} = \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix}
\end{aligned}$$

■

Below, we revisit the derivation of the solution to the linear time-invariant system:

$$\begin{aligned}
&\dot{x}(t) = A\,x(t) + B\,u(t), \qquad x(t_0) = x_0 \\
&y(t) = C\,x(t) + D\,u(t)
\end{aligned}$$

The solution is given by (3.8) and (3.9), as reproduced below:

$$\begin{aligned}
x(t) &= e^{(t-t_0)A}x_0 + \int_{t_0}^{t} e^{(t-\tau)A}B\,u(\tau)\,d\tau \\
y(t) &= C(t)\,e^{(t-t_0)A}x_0 + C(t)\int_{t_0}^{t} e^{(t-\tau)A}B\,u(\tau)\,d\tau + D\,u(t)
\end{aligned}$$

Below, we show that these formulas can be derived via Laplace transform and inverse Laplace transform. Consider the Laplace transform of the linear time-invariant system:

$$\begin{cases} s\hat{X}(s) - x_0 &= A\hat{X}(s) + B\hat{U}(s) \\ s\hat{Y}(s) &= C\hat{X}(s) + D\hat{U}(s) \end{cases}$$

$$\Rightarrow \begin{cases} \hat{X}(s) &= (sI-A)^{-1}x_0 + (sI-A)^{-1}B\hat{U}(s) \\ \hat{Y}(s) &= C(sI-A)^{-1}x_0 + D(sI-A)^{-1}B\hat{U}(s) \end{cases}$$

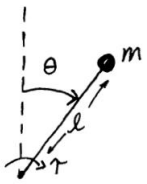
$$\Rightarrow \begin{cases} x(t) &= \mathcal{L}^{-1}\{(sI-A)^{-1}\}x_0 + \mathcal{L}^{-1}\{(sI-A)^{-1}\} \star (Bu)(t) \\ &= e^{(t-t_0)A}x_0 + \int_{t_0}^{t} e^{(t-\tau)A}B\,u(\tau)\,d\tau \\ y(t) &= C\mathcal{L}^{-1}\{(sI-A)^{-1}\}x_0 + D\,\mathcal{L}^{-1}\{(sI-A)^{-1}\} \star (Bu)(t) \\ &= Ce^{(t-t_0)A}x_0 + D\int_{t_0}^{t} e^{(t-\tau)A}B\,u(\tau)\,d\tau \end{cases}$$

Example (**Inverted Pendulum**). Consider the **inverted pendulum** described by the differential equation and figure provided below:

$$ml^2\ddot{\theta} - mgl\sin\theta = \tau$$

Analyze its dynamics.

Figure 10.1



Solution:

The given dynamics can be rewritten as:

$$\ddot{\theta} = \Omega^2\theta + u$$
$$y = \theta$$

where we have defined:

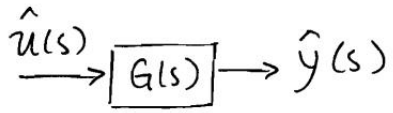
$$\Omega^2 = \frac{g}{l}, \qquad u(t) = \frac{\tau}{ml^2}$$

By taking the Laplace transform on both sides, we have for the transfer function $G(s)$:

$$G(s) \equiv \frac{\hat{Y}(s)}{\hat{U}(s)} = \frac{1}{s^2 - \Omega^2} = \frac{1}{(s+\Omega)(s-\Omega)}$$

The zero-state response of the system can thus be represented by the figure below. This system has an unstable pole at $s = \Omega$.

Figure 10.2



Closed-Loop Solution:

Suppose we try to cancel this pole by passing the input through another controller with transfer function and applying some sort of closed-loop feedback:

$$K(s) = \frac{s - \Omega}{s}$$

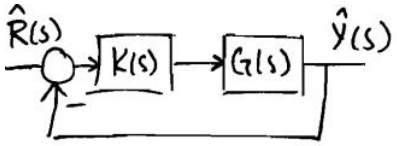
such that the original transfer function, $G(s)$ becomes (see accompanying figure):

$$H(s) \equiv K(s) \cdot G(s) = \frac{1}{s(s + \Omega)}$$

Figure 10.3

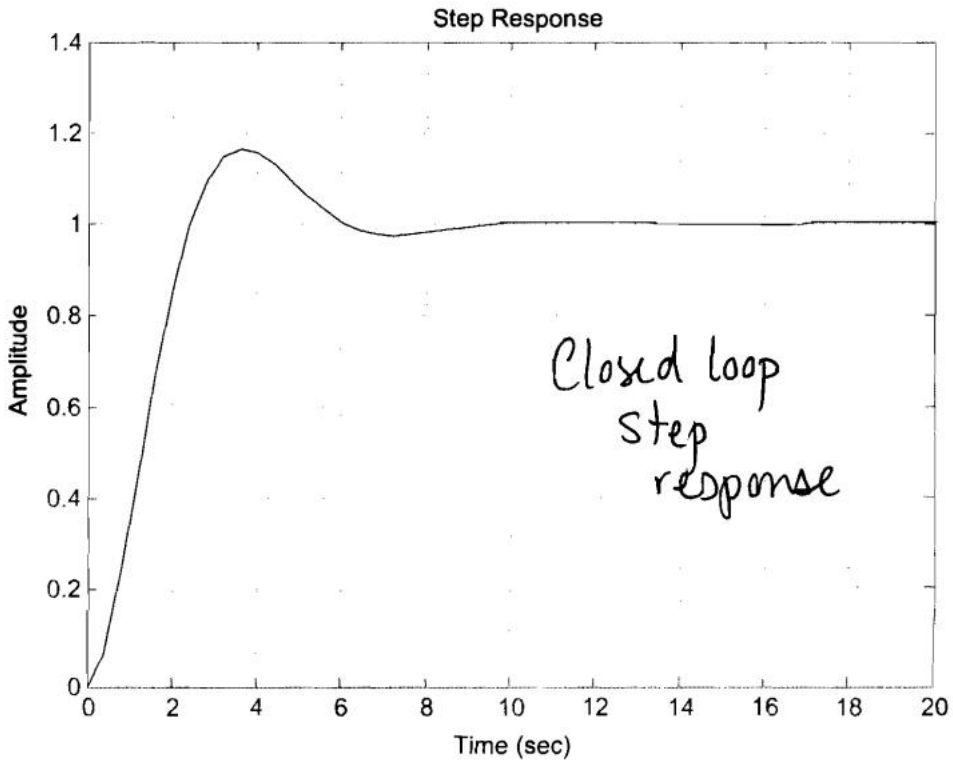
Now, suppose we close the loop:

Figure 10.4



If we apply a step function for the input $R(t)$, the resulting closed-loop step response is as follows (the MATLAB code used to generate figures in this example is provided at the end of the section):

Figure 10.5



Now, Consider the original dynamics as:

$$\dot{x}_1 = x_2$$
$$\dot{x}_2 = \Omega^2 x_1 + u$$

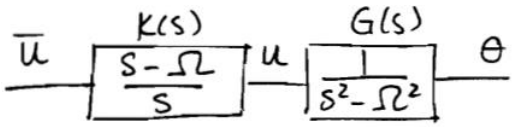
We wish to replace $\bar{u}$ with as the overall input of the composite system. This can be done by introducing a third state variable, x_3, from the relationship between $\bar{u}$ and u:

$$u = \left(\frac{s-\Omega}{s}\right)\bar{u}$$
$$\Rightarrow \dot{u} = \dot{\bar{u}} - \Omega\bar{u}$$

A good candidate for x_3 is thus $x_3 \equiv \bar{u} - u$, which yields $\dot{x}_3 = \Omega\bar{u}$. The state space equations for the above (open loop) system are (see accompanying figure):

$$\dot{x}_1 = x_2$$
$$\dot{x}_2 = \Omega^2 x_1 + \bar{u}$$
$$\dot{x}_3 = \Omega\bar{u}$$

Figure 10.6



MATLAB Simulations for Closed-Loop Step Response and Initial State Response:

For the remainder of this example, we will suppose $\Omega = 1$ for simplicity. If so, the above equations can be rewritten in matrix form as follows:

$$\dot{x} = \underbrace{\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & -1 \\ 0 & 0 & 0 \end{bmatrix}}_{\equiv A} x + \underbrace{\begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}}_{\equiv B} \bar{u}$$
$$y = \underbrace{\begin{bmatrix} 1 & 0 & 0 \end{bmatrix}}_{\equiv C} x$$

Now, consider the closed-loop system obtained by using as our input $\bar{u}$ the difference between a desired output r and the output $y(t)$ (via feedback), as shown below. Here, we take:

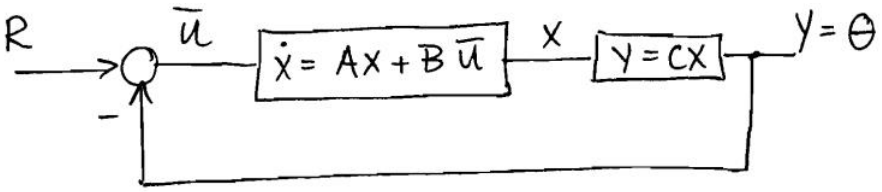
$$u(t) = c_0 \cdot u_{st}(t)$$

where $u_{st}(t)$ denotes the *unit-step function*.

The dynamics of the closed-loop system can be analyzed in the frequency domain as follows (see accompanying figure):

$$\begin{aligned} & u(t) = e(t) = r(t) - y(t) \\ \Rightarrow & H(s) \cdot [R(s) - Y(s)] = Y(s) \\ \Rightarrow & Y(s) = \frac{H(s)}{1 + H(s)} R(s) = \frac{1}{s(s+\Omega) + 1} \cdot \frac{c_0}{s} \end{aligned}$$

Figure 10.8



As a sanity check, we can use the Final-Value Theorem to verify that the steady-state response of the system is stable:

$$\lim_{t \to \infty} y(t) = \lim_{s \to 0^+} sY(s) = c_0$$

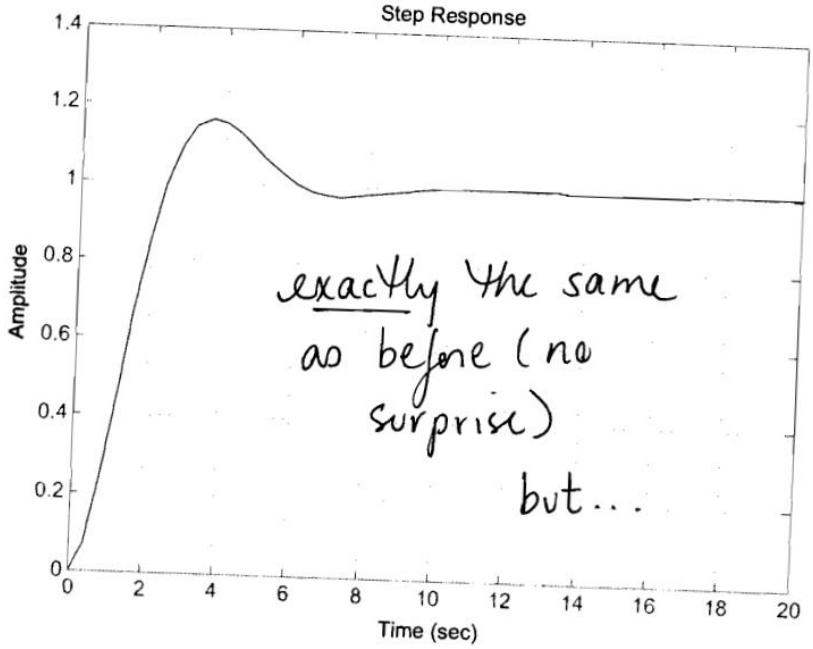
Back in the state space representation, the closed-loop system becomes:

$$\begin{aligned} \dot{x} &= (A - BC)x + BR \\ y &= Cx \end{aligned} \tag{3.10}$$

Consider the following two types of system response simulated and plotted using MATLAB:

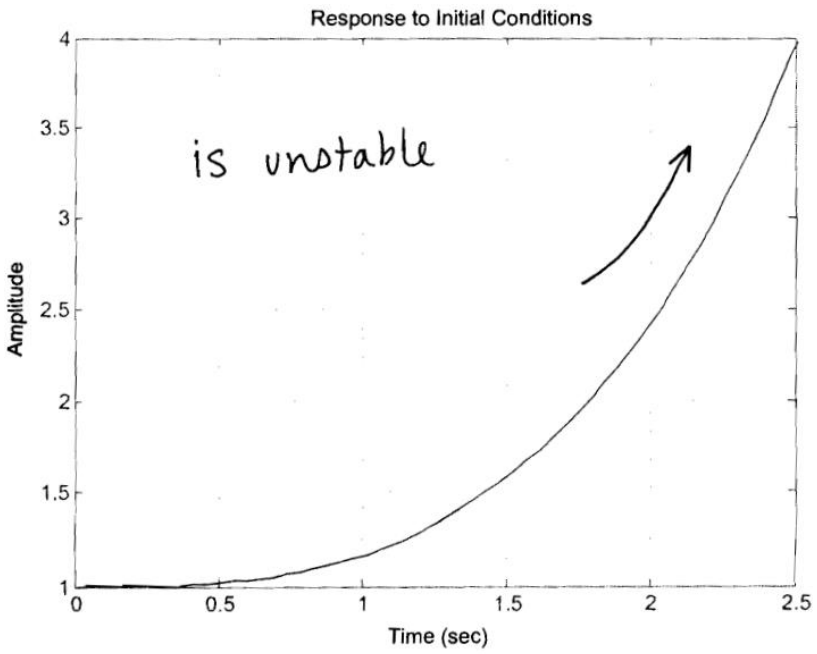
1. Closed-Loop Step Response:

   Figure 10.9



2. Closed-Loop Initial-State Response:

   Figure 10.10



Qualitative Analysis to Explain Instability:

The above simulation results illustrate that, although pole cancellation modifies the poles in the transfer function, which connects the input and output of a system, it cannot eliminate sources of instability within the system itself. Thus, when the system may become unstable when initialized at particular states, even in the absence of external inputs. Below, we qualitatively analyze the state space model to gain more insight into this phenomenon.

By (3.5), the solution to (3.10) is:

$$x(t) = e^{t(A-BC)}x_0 + \int_0^t e^{(t-\tau)(A-BC)} B\, R(\tau)\, d\tau,$$

where $e^{t(A-BC)}$ can be calculated via the inverse Laplace transform:

$$\begin{aligned} e^{t(A-BC)} &= \mathcal{L}^{-1}\{(sI-A)^{-1}\} = \mathcal{L}^{-1}\left\{ \begin{bmatrix} s & -1 & 0 \\ 0 & s & 1 \\ 1 & 0 & s \end{bmatrix}^{-1} \right\} \\ &= \mathcal{L}^{-1}\left\{ \frac{1}{s^3-1} \begin{bmatrix} s^2 & s & -1 \\ 1 & s^2 & -s \\ -s & -1 & s^2 \end{bmatrix}^{-1} \right\} \end{aligned}$$

Note that $(sI-(A-BC))^{-1}$ can be evaluated via Gauss-Jordan elimination, use of the adjunct (classical adjoint) matrix, or any other method for evaluating the inverse of an invertible matrix. In particular, Gauss-Jordan elimination provides an straightforward method for evaluating $(sI-(A-BC))^{-1}$, while the adjunct matrix method reveals that $(sI-(A-BC))^{-1}$

is a rational function with polynomials of degrees $n-1$ and n in the numerator and denominator, respectively.

Recall that we wish to understand why the closed-loop step response of the system is stable, while the initial state response, with $x(t) = (1,0,0)^T$ is not (to be more precise, it is the first term in the response, which is a 3D vector, that is unstable). We do so explicitly below. This time, consider from a qualitative viewpoint the response $y(t)$ under the two types of responses simulated earlier:

1. Closed-Loop Step Response—$R(t) = 0$, and $x_0 = (1,0,0)^T$:

   Similarly, only certain components of $e^{t(A-BC)}$ are necessary in the evaluation of $y(t)$ in this case:

$$\begin{aligned}
y(t) &= \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} x(t) \\
&= \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \left[\int_0^t e^{(t-\tau)A} B \, d\tau \right] \\
&= \int_0^t \left(\begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \left[e^{(t-\tau)A} \right] \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} \right) d\tau \\
&= \int_0^t \mathcal{L}^{-1} \left\{ \frac{s}{s^3 - 1} - \frac{1}{s^3 - 1} \right\} d\tau \\
&= \int_0^t \mathcal{L}^{-1} \left\{ \frac{1}{s^2 + s + 1} \right\} d\tau \\
&= \int_0^t e^{-0.5(t-\tau)} \cos\big(0.866(t-\tau)\big) \, d\tau
\end{aligned}$$

   which is stable, since the poles in this case, $s = -0.5 \pm j\,0.866$ are all on the left half of the complex plane.

2. Closed-Loop Initial-State Response—$R(t) = 0$, and $x_0 = (1,0,0)^T$:

   It is, in fact, unnecessary to evaluate each term in $e^{t(A-BC)}$; the figures given above imply we only have to show that certain parts of the response converge:

$$\begin{aligned}
y(t) &= \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} x(t) \\
&= \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} e^{t(A-BC)} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \\
&= L^{-1} \left\{ \frac{s^2}{s^3 - 1} \right\}
\end{aligned}$$

Since $s^3 - 1$ includes $s = 1$ as a pole, the resulting $y(t)$ is a linear combination of exponential and sinusoidal terms, one of which is e^t. This demonstrates that $y(t)$ has an unstable mode due to the pole at $s = 1$.

For this purpose, it is actually unnecessary to evaluate every term in $e^{t(A-BC)}$; we only need to obtain the $(1, 1)$ entry.

This result can be interpreted by comparing A and $A - BC$:

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & -1 \\ 0 & 0 & 0 \end{bmatrix}$$
$$A - BC = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & -1 \\ 1 & 0 & 0 \end{bmatrix}$$

We can straightforwardly compute the eigenvalues of A and $A - BC$ as follows:

$$\sigma(A) = \{1, -1, 0\}$$
$$\sigma(A - BC) = \left\{1, \frac{1}{2} \pm j\,0.87\right\}$$

Note that the unstable eigenvalue at $s = 1$ is still present, even after the control is applied.

■

Note. The MATLAB code used to generate the figures in this example is as follows:

Figure 10.7

```
% EECS221A CJT
% Inverted Pendulum
% Unstable mode "invisible" in TF but "visible" in state space model

% First define the transfer function and
% plot the closed loop step response
s = tf('s');
G = 1/(s^2-1);
K = (s-1)/s;

% unity feedback closed loop system
step(K*G/(1+K*G));

% Next, define the state space model

A = [0 1 0; 1 0 -1; 0 0 0];
B = [0 1 1]';
C = [1 0 0];
D = 0;

% open loop state space model
sys_ol = ss(A,B,C,D);

% unity feedback closed loop system
sys_cl = ss(A-B*C,B,C,D);

% closed loop step response
step(sys_cl,20);

% initial state response of closed loop system
%initial(sys_cl,[1 0 0]);
```

$\Omega = 1$

3.8 Lecture 10 Discussion

Example (**Discussion 6, Problem 3**). Compute e^{tA} for the following matrix:

$$\begin{bmatrix} -1 & 1 \\ -2 & -3 \end{bmatrix}$$

Below, we compute e^{tA} in three different ways—(1) Laplace Transform, (2) Cayley-Hamilton Theorem, (2) Diagonalization.

Solution 1 : (Laplace Transform)

We have:

$$\begin{aligned} e^{tA} &= \mathcal{L}^{-1}\{(sI - A)^{-1}\} \\ &= \mathcal{L}^{-1}\left\{ \begin{bmatrix} s+1 & -1 \\ 2 & s+3 \end{bmatrix}^{-1} \right\} \\ &= \mathcal{L}^{-1}\left\{ \frac{1}{(s+1)(s+3)+2} \begin{bmatrix} s+3 & 2 \\ 1 & s+1 \end{bmatrix} \right\} \\ &= \mathcal{L}^{-1}\left\{ \frac{1}{(s+2)^2+1} \begin{bmatrix} s+3 & 2 \\ 1 & s+1 \end{bmatrix} \right\} \\ &= \begin{bmatrix} e^{-2t}(\cos t + \sin t) & e^{-2t}\sin t \\ -2e^{-2t}\sin t & e^{-2t}(\cos t - \sin t) \end{bmatrix} \end{aligned}$$

Solution 2 : (Cayley-Hamilton Theorem)

To invoke the Cayley-Hamilton Theorem. which roughly states that all square matrices satisfy their own characteristic equation, we must find the characteristic function of A:

$$\chi_\lambda(A) = (\lambda+1)(\lambda+3)+2 = \lambda^2 + 4\lambda + 5$$

The roots of $\chi_\lambda(A)$, i.e. the eigenvalues of A, are $-2 \pm i$.

Thus, the Cayley-Hamilton Theorem implies that $A^2 + 4A + 5 = 0$. Now, consider the function $e^{t\lambda}$, which we will treat as an infinite polynomial of λ, with coefficients dependent on t. We hope to express $e^{t\lambda}$ in such a way that allows us to easily compute e^{tA} by replacing λ with A. To that end, consider the long division of $e^{t\lambda}$ over $\chi_\lambda(A)$. By the Polynomial Remainder Theorem, there must exist polynomials $q(\lambda), r(\lambda)$, with the degree of $r(\lambda)$ less than 2 (the degree of $\chi_\lambda(A)$, such that:

$$e^{t\lambda} = q(\lambda)\cdot(\lambda^2 + 4\lambda + 5) + \underbrace{(c_1\lambda + c_0)}_{\equiv r(\lambda)}.$$

The coefficients $c_0, c_1 \in \mathbb{C}$ can be readily computed by taking λ to be the eigenvalues of A, since they are the roots of the polynomial $\lambda^2 + 4\lambda + 5$:

$$\begin{aligned} e^{t(-2+i)} &= c_1(-2+i) + c_0 \\ e^{t(-2-i)} &= c_1(-2-i) + c_0 \end{aligned}$$

Solving for c_0, c_1, we have:

$$\begin{aligned} c_1 &= \frac{1}{i2}(e^{it} - e^{-it}) = e^{-2t}\sin t \\ c_2 &= e^{-2t}(e^{it} - \sin t(-2+i)) = e^{-2t}(\cos t + 2\sin t) \end{aligned}$$

Substituting back into e^{tx}, we have:

$$e^{t\lambda} = q(\lambda)\cdot(\lambda^2 + 4\lambda + 5) + (e^{-2t}\sin t)\lambda + e^{-2t}(\cos t + 2\sin t)$$

Finally, replacing λ with A, we have:

$$\begin{aligned} e^{tA} &= q(A)\cdot(A^2 + 4A + 5) + (e^{-2t}\sin t)A + e^{-2t}(\cos t + 2\sin t)I_2 \\ &= (e^{-2t}\sin t)A + e^{-2t}(\cos t + 2\sin t)I_2 \\ &= \begin{bmatrix} e^{-2t}(\cos t + \sin t) & e^{-2t}\sin t \\ -2e^{-2t}\sin t & e^{-2t}(\cos t - \sin t) \end{bmatrix} \end{aligned}$$

Solution 3 : (Diagonalization)

Finally, e^{tA} can also be computed by diagonalizing A, since, if $A = PDP^{-1}$, where P is invertible and D is diagnol, then $e^{tA} = Pe^{tD}P^{-1}$.

From above, we find that the characteristic equation of A is $\chi_\lambda(A) = \lambda^2 + 4\lambda + 5$, with roots at $-2 \pm i$. Using this fact, we find that the following serves as possible values for P and D:

$$P = \begin{bmatrix} 1 & 1 \\ -1+i & -1-i \end{bmatrix}, \qquad D = \begin{bmatrix} -2+i & 0 \\ 0 & -2-i \end{bmatrix}, \qquad P^{-1} = \frac{1}{2}\begin{bmatrix} 1-i & -i \\ 1+i & i \end{bmatrix}$$

Thus, we have:

$$\begin{aligned} e^{tA} &= Pe^{tD}P^{-1} \\ &= \begin{bmatrix} 1 & 1 \\ -1+i & -1-i \end{bmatrix}\begin{bmatrix} e^{-2t}(\cos t + i\sin t) & 0 \\ 0 & e^{-2t}(\cos t - i\sin t) \end{bmatrix}\cdot\frac{1}{2}\begin{bmatrix} 1-i & -i \\ 1+i & i \end{bmatrix} \\ &= \begin{bmatrix} e^{-2t}(\cos t + \sin t) & e^{-2t}\sin t \\ -2e^{-2t}\sin t & e^{-2t}(\cos t - \sin t) \end{bmatrix} \end{aligned}$$

The last equality can be verified by rearranging terms.

Example (**Discussion 6, Problem 6**). Calculate the state transition matrix for $\dot{x}(t) = A(t)x(t)$, where:

$$A(t) = \begin{bmatrix} t & t \\ 0 & -1 \end{bmatrix}$$

Let $x(t_0) = (a, b)^T$, and rewrite the differential equations as:

$$\begin{aligned} \dot{x}_1 &= tx_1 + tx_2, \quad & x_1(t_0) &= a, \\ \dot{x}_2 &= -x_2, & x_2(t_0) &= b, \end{aligned}$$

The second differential equation, which only concerns x_2, yields the solution $x_2(t) = b \cdot e^{-(t-t_0)}$. Thus, the first equation becomes:

$$\begin{aligned}
&\dot{x}_1 - tx_1 = bte^{-(t-t_0)}, \quad x_1(t_0) = a \\
\Rightarrow &\frac{d}{dt}\left(e^{-\frac{1}{2}t^2}x_1\right) = bte^{-\frac{1}{2}t^2-t+t_0} \\
\Rightarrow &e^{-\frac{1}{2}t^2}x_1 - e^{-\frac{1}{2}t_0^2}a = b \cdot e^{t_0}\int_{t_0}^{t} \tau e^{-\frac{1}{2}\tau^2-\tau} \\
\Rightarrow &x_1(t) = a \cdot e^{-\frac{1}{2}(t^2-t_0^2)} + b \cdot e^{\frac{1}{2}t^2+t_0}\int_{t_0}^{t} \tau e^{-\frac{1}{2}\tau^2-\tau}
\end{aligned}$$

where the integrating factor was calculated as $I(t) = e^{\int -\tau d\tau} = e^{-\frac{1}{2}t^2}$ Thus, we have:

$$\begin{aligned}
x(t) &= \begin{bmatrix} a \cdot e^{-\frac{1}{2}(t^2-t_0^2)} + b \cdot e^{\frac{1}{2}t^2+t_0}\int_{t_0}^{t} \tau e^{-\frac{1}{2}\tau^2-\tau} \\ b \cdot e^{-(t-t_0)} \end{bmatrix} \equiv \Phi(t,t_0)\,x(t_0) \\
\Rightarrow \Phi(t,t_0) &= \begin{bmatrix} e^{\frac{1}{2}(t^2-t_0^2)} & e^{\frac{1}{2}t^2+t_0}\int_{t_0}^{t} \tau e^{-\frac{1}{2}\tau^2-\tau} \\ 0 & e^{-(t-t_0)} \end{bmatrix}
\end{aligned}$$

Alternatively, $\Phi(t,t_0)$ could have been derived by directly solving the differential equation:

$$\dot{\Phi}(t,t_0) = A(t)\Phi(t,t_0), \qquad \Phi(t_0,t_0) = I_2$$

and evaluating differential equations for each of the four elements in $\Phi(t,t_0) \in \mathbb{R}^{2\times 2}$. (The same procedure is outlined as an alternate solution in Discussion 5, Problem 8). This is left to the reader as an exercise. In particular, $\Phi_{12}(t)$ (which, as the above derivation shows, is the most "complicated" element in $\Phi(t,t_0)$) satisfies the differential equation:

$$\dot{\Phi}_{12} = t\Phi_{12} + te^{-(t-t_0)}, \qquad \Phi_{12}(t_0) = 0$$

Example (**Discussion 6, Problem 7**). Compute e^{tA_1} and e^{tA_2}, where:

$$A_1 = \begin{bmatrix} 9 & 1 \\ -4 & -5 \end{bmatrix}, \qquad A_2 = \begin{bmatrix} 5 & -3 & 2 \\ 15 & -9 & 6 \\ 10 & -6 & 4 \end{bmatrix}$$

Solution :

1. We will solve this problem using Laplace transform and the Cayley-Hamilton Theorem. Note that diagonalizing A_1 (or finding its Jordan canonical form) involves finding the eigenvectors (and/or generalized eigenvectors) of A, which is a time-consuming process.)

   (a) Laplace Transform:

   Using Laplace transform and inverse Laplace transform, we have:

$$\begin{aligned}
e^{tA_1} &= \mathcal{L}^{-1}\{(sI - A)^{-1}\} \\
&= \mathcal{L}^{-1}\left\{\begin{bmatrix} s+9 & -1 \\ 4 & s+5 \end{bmatrix}\right\} \\
&= \mathcal{L}^{-1}\left\{\frac{1}{(s+9)(s+5)+4}\begin{bmatrix} s+5 & 1 \\ -4 & s+9 \end{bmatrix}\right\} \\
&= \mathcal{L}^{-1}\left\{\frac{1}{(s+7)^2}\begin{bmatrix} s+5 & 1 \\ -4 & s+9 \end{bmatrix}\right\} \\
&= \begin{bmatrix} (1-2t)e^{-7t} & te^{-7t} \\ -4te^{-7t} & (1+2t)e^{-7t} \end{bmatrix}
\end{aligned}$$

(b) Cayley-Hamilton Theorem:

The characteristic equation of A_1 is:

$$\chi_A(\lambda) = (\lambda+9)(\lambda+5)+4 = (\lambda+7)^2$$

Let $q(\lambda)$ and $c_1, c_0 \in \mathbb{R}$ be given such that:

$$\begin{aligned}
e^{t\lambda} &= q(\lambda)\cdot(\lambda+7)^{-1} + c_1\lambda + c_0 \\
te^{t\lambda} &= q'(\lambda)\cdot(\lambda+7)^2 - +2q(\lambda)\cdot(\lambda+7) + c_1,
\end{aligned}$$

where the second equality follows by taking the derivative with respect to λ. Substituting $\lambda = 7$ into the above equation, we have:

$$\begin{aligned}
e^{-7t} &= -7c_1 + c_0 \\
te^{-7t} &= c_1,
\end{aligned}$$

so $(c_1, c_0) = (te^{-7t}, (1+7t)e^{-7t})$. Substituting back into $e^{t\lambda}$, and replacing λ with A, we find:

$$\begin{aligned}
e^{t\lambda} &= q(\lambda)\cdot(\lambda+7)^2 + te^{-7t}\cdot\lambda + (1+7t)e^{-7t} \\
e^{tA} &= q(A)\cdot(A+7I_2)^2 + te^{-7t}\cdot A + (1+7t)e^{-7t}I_2 \\
&= te^{-7t}\cdot A + (1+7t)e^{-7t}I_2 \\
&= \begin{bmatrix} (1-2t)e^{-7t} & te^{-7t} \\ -4te^{-7t} & (1+2t)e^{-7t} \end{bmatrix}
\end{aligned}$$

2. We will solve this problem using Jordan decomposition and the Cayley-Hamilton Theorem. As we will see below, the Jordan decomposition approach actually allows us to find e^{tA} without actually finding the eigenvectors and generalized eigenvectors of A. Since $A \in \mathbb{R}^{3\times3}$, Laplace transform in fact becomes a more time-consuming approach for this particular problem.

(a) Jordan Decomposition:

By observation, the three columns of A_2 are simply multiples of the same vector, $(1, 3, 2)^T$. Thus, two of the three eigenvalues of A_2 are 0, with two linearly independent eigenvectors, $(2, 0, -5)^T$ and $(0, 2, 3)^T$. Since $tr(A_2) = 0$ is the sum of the three eigenvalues, the third eigenvalue must also be 0. If there existed a third linearly independent eigenvector for A_2, then A_2 is diagonalizable with diagonal matrix $D = O$, which would in turn imply $A_2 = O$, a contradiction. This implies that A_2 can only be Jordan decomposed, i.e. $A_2 = PJP^{-1}$, where:

$$J = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Note that $J^2 = O$. Thus, by direct expansion, we have:

$$\begin{aligned} e^{tA_2} &= Qe^{tJ}Q^{-1} = Q(1 + tJ)Q^{-1} = I + tA_2 \\ &= \begin{bmatrix} 1 + 5t & -3t & 2t \\ 15t & 1 - 9t & 6t \\ 10t & -6t & 1 + 4t \end{bmatrix} \end{aligned}$$

(b) Cayley-Hamilton Theorem:

Again, by observation, we note that the three eigenvalues of A_2 are all 0. Thus, $\chi_A(\lambda) = \lambda^3$. Let:

$$\begin{aligned} e^{t\lambda} &= q(\lambda) \cdot \lambda^3 + c_2\lambda^2 + c_1\lambda + c_0 \\ te^{t\lambda} &= q_2(\lambda) \cdot \lambda^3 + 2c_2\lambda + c_1 \\ t^2e^{t\lambda} &= q_3(\lambda) \cdot \lambda^3 + 2c_2 \end{aligned}$$

where the last two equalities were derived by differentiating the first once and twice, respectively, and $q_2(\lambda)$ and $q(\lambda)$ represent polynomials of λ that depend on $q(\lambda)$. Substituting $\lambda = 0$, we find that:

$$(c_2, c_1, c_0) = \left(\frac{1}{2}t^2, t, 1\right).$$

Substituting back into $e^{t\lambda}$, and replacing λ with A_2, we have:

$$\begin{aligned} e^{t\lambda} &= q(\lambda) \cdot \lambda^3 + \frac{1}{2}t^2 \cdot \lambda^2 + t \cdot \lambda + 1 \\ \Rightarrow e^{tA_2} &= \frac{1}{2}t^2A_2^2 + tA_2 + I_3 = tA_2 + I_3, \end{aligned}$$

where $A_2^2 = O$ (it takes some time to verify this).

Chapter 4

System Stability

4.1 Lecture 12

Consider the time-invariant dynamical system:

$$\dot{x}(t) = A(t)\,x(t),\, x(0) = x_0$$

which, as shown in Lecture 10, has the state transition matrix:

$$\Phi(t, 0) = e^{tA}$$

The *Laplace Transform* method of evaluating the matrix polynomial e^{tA} is provided below.

$$\Phi(t, 0) = e^{tA} = \mathcal{L}^{-1}\{(sI - A)^{-1}\}.$$

The expression:

$$(sI - A)^{-1} = \frac{\text{adj}(sI - A)}{\det(sI - A)},$$

where $\det(sI - A)$ and $\text{adj}(sI - A)$ (classical adjoint) are n-dimensional and $n-1$-dimensional matrices, respectively, can be used to evaluate $(sI - A)^{-1}$. Alternatively, $(sI - A)^{-1}$ can also be evaluated via Gauss-Jordan elimination.

Below, we introduce a very important property of the characteristic equation of a square matrix. However, before doing so, we must introduce the concept of a matrix function.

Theorem 4.1 (**Cayley-Hamilton Theorem**). *Let* $\mathrm{A} \in \mathbb{R}^{n\times n}$*, and suppose its characteristic polynomial has the form:*

$$\chi_{\mathrm{A}}(s) \equiv \det(sI - \mathrm{A}) = s^n + d_1 s^{n-1} + \cdots + d_{n-1}s + d_n$$

Then:

$$\chi_A(A) = A^n + d_1 A^{n-1} + \cdots + d_{n-1}A + d_n I = O$$

Remark. In the statement of the Cayley-Hamilton Theorem, we have overloaded the notation $\chi_A(\cdot)$ to describe polynomials of either scalar or square matrix inputs with the same coefficients as the characteristic polynomial of A.

Proof. Consider the inverse of $(sI - A)$, computed using the classical adjoint of A:

$$\begin{aligned}(sI - A)^{-1} &= \frac{\text{adj}(sI - A)}{\det(sI - A)} \\ \Rightarrow (sI - A) \cdot \text{adj}(sI - A) &= \det(sI - A) \cdot I\end{aligned}$$

Let $\text{adj}(A)$, which is of degree $n - 1$, have the following form:

$$\text{adj}(A) = B_0 s^{n-1} + B_1 s^{n-2} + \cdots + B_{n-2} s + B_{n-1}$$

Then the above equation becomes:

$$\begin{aligned}&(sI - A)(B_0 s^{n-1} + B_1 s^{n-2} + \cdots + B_{n-2} s + B_{n-1}) \\ =&(s^n + d_1 s^{n-1} + \cdots + d_{n-1} s + d_n) I\end{aligned}$$

By comparing coefficients on both sides of the equation, we have:

$$\begin{aligned}B_0 &= I \\ B_k &= AB_{k-1} + d_k I, \qquad \forall k = 1, \cdots, n-1 \\ O &= AB_{n-1} + d_n I\end{aligned}$$

We claim that, for each $k \in 1, \cdots, n$:

$$O = A^k B_{n-k} + A^{k-1} d_{n-k+1} + \cdots + d_{n-1} A + d_n$$

This can be done via induction, by working backwards from $n - 1$. When $k = 1$, we have $O = AB_{n-1} + d_{n-1} I$. Suppose the claim holds for some $k \in \{1, \cdots, n-1\}$. Then:

$$\begin{aligned}O &= A^k B_{n-k} + A^{k-1} d_{n-k+1} + \cdots + d_{n-1} A + d_n \\ &= A^k (AB_{n-k-1} + d_{n-k}) + A^{k-1} d_{n-k+1} + \cdots + d_{n-1} A + d_n \\ &= A^{k+1} B_{n-k-1} + A^k d_{n-k} + A^{k-1} B_{n-k+1} + \cdots + d_{n-1} A + d_n\end{aligned}$$

By induction, the claim holds. Taking $k = n - 1$ completes the proof. ■

The Cayley-Hamilton Theorem is both of significant theoretical interest in linear and abstract algebra, and useful as a computational tool for evaluating functions of square matrices. Alternative proofs for the Cayley-Hamilton (some of which may appear less ad hoc than the one presented above) are provided in the appendix.

Now, consider a polynomial $p(s)$ of degree $m \geq n$ (where m may be infinite), and suppose we wish to evaluate the matrix $p(A)$. By the Polynomial Remainder Theorem, there exists polynomials $q(s)$ and $r(s)$ such that:

$$p(s) = q(s) \cdot \chi_A(s) + r(s),$$

where the degree of $q(s)$ is $m - n$, if m is finite, and infinite otherwise, and the degree of $r(s)$ is less than n. By the Cayley-Hamilton Theorem, $\chi_A(A) = 0$, so:

$$p(A) = q(A) \cdot \chi_A(A) + r(A) = r(A)$$

We have thus reduced the problem of evaluating a matrix polynomial of $\geq n$ (and possibly infinite) degree to one of solving a matrix polynomial of $< n$ degree. This is particularly useful when $n = \infty$, i.e. when $p(s)$ is an infinite polynomial, as is the case when $p(s)$ represents the Taylor expansion of a continuously differentiable, non-polynomial function (i.e. $\sin s, \cos s, e^s$, etc.) Examples are given in the Discussion notes following this lecture.

Dyadic Expansion:

We devote the next part of this section to a discussion of *dyadic expansion*. Let $A \in \mathbb{R}^{n\times n}$ be a diagonalizable matrix with n distinct eigenvalues $\lambda_1, \lambda_2 \cdots, \lambda_n$, and corresponding *right (column) eigenvectors* $e_1, e_2 \cdots, e_n$. Define:

$$P \equiv \begin{bmatrix} e_1 & e_2 & \cdots & e_n \end{bmatrix},$$
$$D \equiv \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}.$$

Then we have:

$$A \underbrace{\begin{bmatrix} e_1 & e_2 & \cdots & e_n \end{bmatrix}}_{\equiv P} = \begin{bmatrix} \lambda_1 e_1 & \lambda_2 e_2 & \cdots & \lambda_n e_n \end{bmatrix}$$
$$= \underbrace{\begin{bmatrix} e_1 & e_2 & \cdots & e_n \end{bmatrix}}_{\equiv P} \underbrace{\begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}}_{\equiv D}$$
$$\Rightarrow A = PDP^{-1}$$

The above process is known as the *diagonalization* of A.

Now, let $v_1^T, v_2^T, \cdots, v_n^T$ be <u>the rows of P^{-1}</u>. Then:

$$A = PDP^{-1} = (P^{-1})^{-1}DP$$
$$= \begin{bmatrix} v_1^T \\ v_2^T \\ \vdots \\ v_n^T \end{bmatrix}^{-1} \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} v_1^T \\ v_2^T \\ \vdots \\ v_n^T \end{bmatrix}$$
$$\Rightarrow \begin{bmatrix} v_1^T \\ v_2^T \\ \vdots \\ v_n^T \end{bmatrix} A = \begin{bmatrix} v_1^T \\ v_2^T \\ \vdots \\ v_n^T \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} = \begin{bmatrix} \lambda_1 v_1^T \\ \lambda_2 v_2^T \\ \vdots \\ \lambda_n v_n^T \end{bmatrix}$$

It follows that $v_i^T A = \lambda_i v_i^T$, for each $i = 1, \cdots, k$. We call $\{v_1^T, \cdots, v_n^T\}$ the *left (row) eigenvectors* of A.

Now, let us return to the expression $A = PDP^{-1}$, and rewrite P and P^{-1} as a collection of the column and row eigenvectors of A, respectively:

$$\begin{aligned} A &= PDP^{-1} \\ &= \begin{bmatrix} e_1 & e_2 & \cdots & e_n \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} v_1^T \\ v_2^T \\ \vdots \\ v_n^T \end{bmatrix} \\ &= \sum_{i=1}^{n} \lambda_i e_i v_i^T, \end{aligned}$$

where each $e_i v_i^T$ is a rank-1 matrix called a *dyad*.

Observe that:

$$I = PP^{-1} = \begin{bmatrix} e_1 & e_2 & \cdots & e_n \end{bmatrix} \begin{bmatrix} v_1^T \\ v_2^T \\ \vdots \\ v_n^T \end{bmatrix} = \sum_{i=1}^{n} e_i v_i^T$$

$$I = P^{-1}P = \begin{bmatrix} v_1^T \\ v_2^T \\ \vdots \\ v_n^T \end{bmatrix} \begin{bmatrix} e_1 & e_2 & \cdots & e_n \end{bmatrix} = \begin{bmatrix} v_1^T e_1 & v_1^T e_2 & \cdots & v_1^T e_n \\ v_2^T e_1 & v_2^T e_2 & \cdots & v_2^T e_n \\ \vdots & \vdots & \ddots & \vdots \\ v_n^T e_1 & v_n^T e_2 & \cdots & v_n^T e_n \end{bmatrix}$$

The first equality states that the dyads of A sum to the identity matrix, while the second implies that:

$$v_i^T e_j = \delta_{ij},$$

where δ_{ij}, called the *Kronecker delta*, is 1 if $i = j$ and 0 otherwise.

Example. Find the minimal polynomial of:

$$A = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

Solution :

Since $\sigma(A) = \{2\}$, the characteristic polynomial of A is:

$$\chi_A(s) = (s-2)^3$$

However, in fact, we have:

$$(A - 2I)^2 = 0$$

This demonstrates that the characteristic polynomial is not always the minimal polynomial.

Theorem 4.2. *If $A \in \mathbb{C}^{n\times n}$ has n distinct eigenvalues $\lambda_1, \cdots, \lambda_n \in \mathbb{C}$, with corresponding eigenvectors $v_1, \cdots, v_n$, then $\{v_1, \cdots, v_n\}$ forms a basis for $\mathbb{C}^{n\times n}$.*

Proof. Suppose by contradiction that $\{v_1, \cdots, v_n\}$ are not linearly independent. Then there exist $\alpha_1, \cdots, \alpha_n \in \mathbb{C}$, not all zero, such that:

$$\alpha_1 v_1 + \cdots + \alpha_n v_n = 0.$$

Without loss of generality, suppose $\alpha_1 \neq 0$. Applying $(L - \lambda_2 I)(L - \lambda_3 I)\cdots(L - \lambda_n I)$ on both sides, we have:

$$\begin{aligned} 0 &= (L - \lambda_2 I)(L - \lambda_3 I)\cdots(L - \lambda_n I)(\alpha_1 v_1 + \cdots + \alpha_n v_n) \\ &= \alpha_1 (L - \lambda_2 I)(L - \lambda_3 I)\cdots(L - \lambda_n I) v_1 + 0 \\ &= \alpha_1 (\lambda_1 - \lambda_2)(\lambda_1 - \lambda_3)\cdots(\lambda_1 - \lambda_n) v_1 \\ &\neq 0, \end{aligned}$$

a contradiction. The above equalities follow from the facts that (1) The $(L - \lambda_i I)$ terms in the product $(L - \lambda_2 I)(L - \lambda_3 I)\cdots(L - \lambda_n I)$ all commute, since they are all functions of the same mapping L, and (2) $(L - \lambda_i I)v_i = 0$ for each $i = 1, \cdots, n$. ■

Alternatively, this theorem can be proved via induction, as demonstrated below.

Proof. The lemma can be proven using by induction. When $k = 1$, we have $Lv_1 = \lambda_1 v_1$, where $v_1 \neq 0$, so $\{v_1\}$ is a linearly independent subset. Suppose the lemma is valid for some $k - 1$, where $k \in \{2, 3, \cdots\}$. Let:

$$a_1 v_1 + a_2 v_2 + \cdots + a_k v_k = 0$$

Then:

$$\begin{aligned} 0 &= (L - \lambda_k I)(a_1 v_1 + a_2 v_2 + \cdots + a_k v_k) \\ &= a_1(\lambda_1 - \lambda_k)v_1 + a_2(\lambda_2 - \lambda_k)v_2 + \cdots + a_{k-1}(\lambda_{k-1} - \lambda_k)v_{k-1} + \underbrace{a_k(\lambda_k - \lambda_k)v_k}_{=0} \end{aligned}$$

By the induction hypothesis, $\{v_1, v_2, \cdots, v_{k-1}\}$ is linearly independent, so:

$$a_1(\lambda_1 - \lambda_k) = a_2(\lambda_2 - \lambda_k) = \cdots = a_{k-1}(\lambda_{k-1} - \lambda_k) = 0$$

Since $\lambda_i \neq \lambda_k$ for each $i \neq k$:

$$a_1 = a_2 = \cdots = a_{k-1} = 0$$

So Equation (4.1) becomes $a_k v_k = 0$. Since $v_k \neq 0$, we have $a_k = 0$, which indicates that $\{v_1, v_2, \cdots, v_k\}$ is a linearly independent subset. The proof follows by induction. ■

This theorem can be applied to the solving process of a matrix differential equation $\dot{x} = Ax$. Recall that its solution is:

$$x(t) = e^{tA}x_0$$

If A has n distinct eigenvalues $\lambda_1, \cdots, \lambda_n \in \mathbb{C}$, with corresponding eigenvalues $v_1, \cdots, v_n$, then by the above theorem, we can write:

$$x_0 = \sum_{i=1}^{\infty} \eta_i(0) v_i,$$

which allows us to simplify the given expression for $x(t)$:

$$\begin{aligned} x(t) = e^{tA}x_0 &= \sum_{i=1}^{\infty} \eta_i(0) e^{tA} v_i \\ &= \sum_{i=1}^{\infty} \eta_i(0) e^{t\lambda_i} v_i \end{aligned}$$

Note that $\{\lambda_1, \cdots, \lambda_n\}$ may contain real values and/or complex conjugate pairs. In general, a complex eigenvalue $\lambda_{ir} + i\lambda_{ic}$ would correspond to either an exponentially growing mode (if $\lambda_{ir} > 0$) or an exponentially decaying mode (if $\lambda_{ir} < 0$), oscillating at angular frequency $|\lambda_{ic}|$.

Modal Decomposition

Recall that, given a change of coordinates in the system state $\overline{x} = Tx$, the equivalent system representation of an LTI system $R : (A, B, C, D)$ is $\overline{R} : (TAT^{-1}, TB, CT^{-1}, D)$, i.e.:

$$\begin{cases} \dot{x} = Ax + Bu, \\ y = Cx + Du, \end{cases} \qquad \Rightarrow \qquad \begin{cases} \dot{\overline{x}} = TAT^{-1}\overline{x} + TBu, \\ y = CT^{-1}\overline{x} + Du, \end{cases}$$

Equivalent systems have the same transfer function, since:

$$\begin{aligned} & CT^{-1}(sI - TAT^{-1})^{-1}TB \\ =& CT^{-1}\big(T(sI - A)^{-1}T^{-1}\big)TB \\ =& C(sI - A)^{-1}B \end{aligned}$$

Let an LTI system $R : (A, B, C, D)$ be given, and suppose A is diagonalizable. Define the rows of B and the columns of C by:

$$B = \begin{bmatrix} b_1 & b_2 & \cdots b_{n_i} \end{bmatrix}, \qquad C = \begin{bmatrix} c_1^T \\ c_2^T \\ \vdots \\ c_{n_o}^T \end{bmatrix}$$

Let T^{-1} be an invertible matrix whose columns are linearly independent right eigenvectors of A. (Equivalently, the rows of T are the corresponding linearly independent left eigenvectors of A). Then we can diagonalize A as shown below:

$$TAT^{-1} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}$$

Then, if we define:

$$TB = \begin{bmatrix} v_1^T \\ v_2^T \\ \vdots \\ v_n^T \end{bmatrix} \begin{bmatrix} b_1 & b_2 & \cdots & b_{n_i} \end{bmatrix} = \begin{bmatrix} v_1^T b_1 & v_1^T b_2 & \cdots & v_1^T b_{n_i} \\ v_2^T b_1 & v_2^T b_2 & \cdots & v_2^T b_{n_i} \\ \vdots & \vdots & \ddots & \vdots \\ v_n^T b_1 & v_n^T b_2 & \cdots & v_n^T b_{n_i} \end{bmatrix} \equiv \begin{bmatrix} \tilde{B}_1 \\ \tilde{B}_2 \\ \vdots \\ \tilde{B}_n \end{bmatrix},$$

$$CT^{-1} = \begin{bmatrix} c_1^T \\ c_2^T \\ \vdots \\ c_{n_o}^T \end{bmatrix} \begin{bmatrix} e_1 & e_2 & \cdots e_n \end{bmatrix} = \begin{bmatrix} c_1^T e_1 & c_1^T e_2 & \cdots & c_1^T e_n \\ c_2^T e_1 & c_2^T e_2 & \cdots & c_2^T e_n \\ \vdots & \vdots & \ddots & \vdots \\ c_{n_o}^T e_1 & c_{n_o}^T e_2 & \cdots & c_{n_o}^T e_n \end{bmatrix} \equiv \begin{bmatrix} \tilde{C}_1 & \tilde{C}_2 & \cdots & \tilde{C}_n \end{bmatrix}.$$

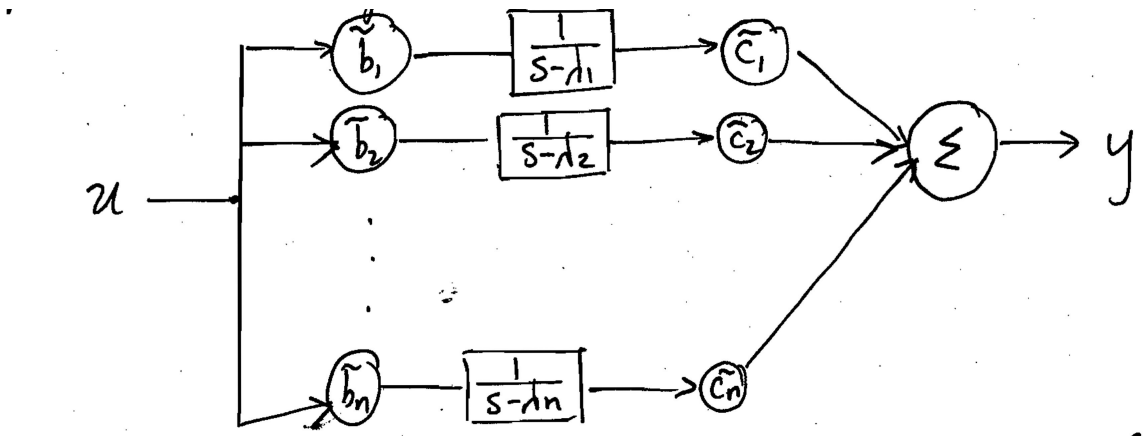
where, as before, $\{e_1, \cdots, e_n\}$ and $\{v_1^T, \cdots, v_n^T\}$ are the left (column) eigenvectors and corresponding left (row) eigenvectors of A, respectively. The transfer function of the system then becomes:

$$\begin{aligned} H(s) &= CT^{-1}(sI - TAT^{-1})^{-1}TB \\ &= \begin{bmatrix} \tilde{C}_1 & \tilde{C}_2 & \cdots & \tilde{C}_n \end{bmatrix} \begin{bmatrix} \frac{1}{s-\lambda_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{s-\lambda_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{s-\lambda_n} \end{bmatrix} \begin{bmatrix} \tilde{B}_1 \\ \tilde{B}_2 \\ \vdots \\ \tilde{B}_n \end{bmatrix} \end{aligned}$$

If the system were *single-input-single-output* (SISO), i.e. $B \in \mathbb{R}^{n\times 1}, C \in \mathbb{R}^{1\times n}$, then each $\tilde{B}_i$ and each $\tilde{C}_i$ is a scalar. The transfer function then has the simple form:

$$H(s) = \sum_{i=1}^{n} \frac{\tilde{C}_i \tilde{B}_i}{s - \lambda_i}$$

This is called **modal decomposition**. Notice that, if $\tilde{B}_i = 0$ or $\tilde{C}_i = 0$, the transfer function $C(sI - A)^{-1}B$ does not contain the corresponding term $1/(s - \lambda_i)$, and thus does not contain the corresponding pole λ_i. A block diagram is given below.



If the system were multiple-input multiple-output (MIMO), the same analysis follows—If, for some $i = 1, \cdots, n$, we have:

$$\tilde{B}_i = \begin{bmatrix} v_i^T b_1 & v_i^T b_2 & \cdots & v_i^T b_n \end{bmatrix} = 0, \quad \text{or}$$

$$\tilde{C}_i = \begin{bmatrix} c_1^T e_i \\ c_2^T e_i \\ \vdots \\ c_n^T e_i \end{bmatrix} = 0,$$

then the pole λ_i does not appear in $C(sI - A)^{-1}B$.

Forced Response

The discussion above implies that, for a given LTI system $R : (A, B, C)$, and arbitrary initial state x_0 and input $u(\cdot)$, exponential functions of the form $e^{\lambda_i t}$, where $\lambda_i \in \sigma(A)$, would arise in the resulting trajectory $x(t)$. We now wish to ask whether there exist specific initial states from which, if the system were excited by an input $u(t) = u_0 e^{\lambda t}$, $\lambda \notin \sigma(A)$, the resulting trajectory would only contain terms of the form $e^{\lambda t}$. The theorem below answers this in the affirmative.

Theorem 4.3. *Given an LTI system $R : (A, B, C)$, and an input $u(t) = u_0 e^{\lambda t}$, with $\lambda \notin \sigma(A)$ and some arbitrary $u_0 \in \mathbb{R}^{n_i}$, the resulting trajectory is of the form:*

$$y(t) = y_0 e^{\lambda t}$$

if and only if the initial condition satisfies:

$$x_0 = (\lambda I - A)^{-1} B u_0$$

Proof. Using (3.9) :

$$\begin{aligned}
\because y(t) &= Ce^{tA}x_0 + \int_0^t Ce^{(t-\tau)A}Bu(\tau)\,d\tau \\
&= Ce^{tA}x_0 + \int_0^t Ce^{(t-\tau)A}Bu_0e^{\lambda\tau}\,d\tau \\
\Rightarrow Y(s) &= C(sI-A)^{-1}x_0 + C(sI-A)^{-1}B\cdot\frac{u_0}{s-\lambda}
\end{aligned}$$

where we have used the fact that the Laplace transform of the convolution of two functions equals the product of the Laplace transform of the individual functions.

Now, we wish to isolate the effect of $(sI-A)^{-1}$ and $(s-\lambda)^{-1}$ on u_0, in order to isolate a term in $Y(s)$ that depends only on $(s-\lambda)^{-1}$. This can be done by rewriting $(sI-A)^{-1}$ in terms of $(\lambda_i - A)^{-1}$ and $(s-\lambda)^{-1}$. In particular, observe that:

$$\begin{aligned}
&(sI-A)-(\lambda I-A) = (s-\lambda)I, \\
\Rightarrow &(\lambda I-A)^{-1} - (sI-A)^{-1} = (s-\lambda)(sI-A)^{-1}(\lambda I-A)^{-1}, \\
\Rightarrow &(sI-A)^{-1} = \big[I-(s-\lambda)(sI-A)^{-1}\big](\lambda I-A)^{-1}
\end{aligned}$$

Substituting into the above expression, we have:

$$\begin{aligned}
Y(s) &= C(sI-A)^{-1}x_0 + C(sI-A)^{-1}B\cdot\frac{u_0}{s-\lambda} \\
&= C(sI-A)^{-1}x_0 + C\big[I-(s-\lambda)(sI-A)^{-1}\big](\lambda I-A)^{-1}B\cdot\frac{u_0}{s-\lambda} \\
&= C(sI-A)^{-1}\big[x_0-(\lambda I-A)^{-1}Bu_0\big] + C(\lambda I-A)^{-1}B\cdot\frac{u_0}{s-\lambda}, \\
\Rightarrow y(t) &= Ce^{tA}\big[x_0-(\lambda I-A)^{-1}Bu_0\big] + C(\lambda I-A)^{-1}Bu_0e^{\lambda t}
\end{aligned}$$

Thus, $y(t)$ is proportional to $e^{\lambda t}$ if and only if $x_0 = (\lambda I-A)^{-1}Bu_0$, in which case:

$$y(t) = C(\lambda I-A)^{-1}Bu_0e^{\lambda t} = H(\lambda)\cdot u_0e^{\lambda t},$$

where $H(s) = C(sI-A)^{-1}B$ is the transfer function of the system. ■

Remark. If the system is stabilizable, then $e^{tA}\to O$ as $t\to\infty$, so $y(t)\to H(\lambda)\cdot u_0e^{\lambda t}$ as $t\to\infty$. In other words, even though $y(t)$ may initially not be directly proportional to $e^{\lambda t}$, it will eventually asymptotically approach a function proportional to $e^{\lambda t}$ as its *steady-state response.*

We end our discussion by connecting the above concepts to that of the "zeros" of a transfer function.

Definition 4.4 (**Zero** (of a transfer function))**.** *Let $R:(A,B,C)$ be a "square" system, in the sense that $n_i = n_0$. Then $z\in\mathbb{C}$ is said to be a zero of the system if the transfer function evaluated at z:*

$$H(z) = C(zI-A)^{-1}B$$

is singular.

For "square" LTI systems with zeros, particular choices of initial state and input functions can be chosen such that the output is identically zero. Specifically, if $H(z)$ is singular for some $z \in \mathbb{C}$, then there must exist some $u_0 \in \mathbb{R}^{n_i}$ such that $H(z)u_0 = 0$. Now, let:

$$\begin{aligned} u(t) &= u_0 e^{zt}, \\ x_0 &= (zI - A)^{-1} B u_0 \end{aligned}$$

The analysis above shows that $y(t) = 0$.

4.2 Lecture 12 Discussion

Example (**Discussion 7, Problem 1**). Let:

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix}$$

Is $\{I_2, A, A^2\}$ linearly dependent or independent in $\mathbb{R}^{2\times 2}$?

Solution :

The characteristic equation of A is:

$$\chi_A(\lambda) = (\lambda - 1)(\lambda - 2) = \lambda^2 - 3\lambda + 2$$

The Cayley-Hamilton Theorem thus implies that:

$$O = \chi_A(A) = (A - I_2)(A - 2I) = A^2 - 3A + 2I_2$$

This shows that $\{I, A, A^2\}$ is linearly dependent.

Remark. The solution to this problem can actually be generalized to show that $\{I, A, \cdots, A^n\}$ is linearly dependent for each $A \in \mathbb{R}^{n\times n}$.

Example (**Discussion 7, Problems 2, 3, 6, 7**). Let:

$$A = \begin{bmatrix} 1 & 0 \\ 1 & 2 \end{bmatrix}$$

Consider the following questions:

1. Find A^3.
2. Find e^{tA} via dyadic expansion.
3. Find $\sin(e^A)$.
4. Find $2A^4 - 3A^3 - 3A^2 + 4I_2$.

Solution :

The characteristic equation of A is:

$$\chi_A(\lambda) = (\lambda - 1)(\lambda - 2) = \lambda^2 - 3\lambda + 2$$

The Cayley-Hamilton Theorem thus implies that:

$$O = \chi_A(A) = (A - I_2)(A - 2I) = A^2 - 3A + 2I_2$$

The three given functions of A can thus be solved via long-division.

1. By long division:

$$A^3 = (A^2 - 3A + 2I_2)(A + 3I_2) + (7A - 6I_2) = 7A - 6I_2 = \begin{bmatrix} 1 & 0 \\ 7 & 8 \end{bmatrix}$$

2. Since we have already computed $\sigma(A) = \{2, 1\}$ to be the eigenvalues of A, the diagonalization of A can be readily found:

$$A = \left[\begin{array}{c|c} 0 & 1 \\ 1 & -1 \end{array}\right] \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \left[\begin{array}{c|c} 1 & 1 \\ \hline 1 & 0 \end{array}\right]$$

where the vertical and horizontal lines are used to emphasize the placement of the (right) column and (left) eigenvectors of A in its diagonalization.

Thus, the dyadic expansion of A is:

$$A = 2 \cdot \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \end{bmatrix} + 1 \cdot \begin{bmatrix} 1 \\ -1 \end{bmatrix} \begin{bmatrix} 1 & 0 \end{bmatrix}$$

Since the eigenvalues of e^{tA} are simply $\{e^{t\lambda_i}\}$, where $\{\lambda_i\}$ are the eigenvalues of A, with the same corresponding eigenvectors, the dyadic expansion of e^{tA} is simply that of A with each eigenvalue λ replaced with e^{tA}:

$$\begin{aligned} A &= e^{2t} \cdot \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \end{bmatrix} + e^t \cdot \begin{bmatrix} 1 \\ -1 \end{bmatrix} \begin{bmatrix} 1 & 0 \end{bmatrix} \\ &= e^{2t} \cdot \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix} + e^t \cdot \begin{bmatrix} 1 & 0 \\ -1 & 0 \end{bmatrix} \\ &= \begin{bmatrix} e^t & 0 \\ e^{2t} - e^t & e^{2t} \end{bmatrix} \end{aligned}$$

3. Let c_1, c_0 be time-dependent coefficients such that:

$$\sin(e^\lambda) = q(\lambda) \cdot (\lambda^2 - 3\lambda + 2) + c_1\lambda + c_0$$

Taking $\lambda = 1$ and $\lambda = 2$, we have:

$$\begin{aligned} \sin(e) &= c_1 + c_0 \\ \sin(e^2) &= 2c_1 + c_0 \end{aligned}$$

Thus, $c_1 = \sin(e^2) - \sin(e)$ and $c_0 = -\sin(e^2) + 2\sin(e)$, so:

$$\begin{aligned} \sin(e^\lambda) &= q(\lambda) \cdot (\lambda^2 - 3\lambda + 2) + (\sin(e^2) - \sin(e))\lambda + (-\sin(e^2) + 2\sin(e)) \\ \Rightarrow \sin(e^A) &= q(A) \cdot (A^2 - 3A + 2I_2) + (\sin(e^2) - \sin(e))A + (-\sin(e^2) + 2\sin(e))I_2 \\ &= (\sin(e^2) - \sin(e))A + (-\sin(e^2) + 2\sin(e))I_2 \\ &= \begin{bmatrix} \sin(e) & 0 \\ \sin(e^2) - \sin(e) & \sin(e^2) \end{bmatrix} \end{aligned}$$

4. By long division:

$$2A^4 - 3A^3 - 3A^2 + 4I_2 = (A^2 - 3A + 2I_2)(2A^2 + 3A + 2I_2) = O$$

i.e. $f(x) = 2x^4 - 3x^3 - 3x^2 + 4$ annihilates A.

Example (**Discussion 7, Problem 4, 9**). Consider:

$$\dot{x} = Ax, \qquad x(0) = v_i$$

where $A \in \mathbb{R}^{n\times n}$ has n linearly independent eigenvectors $v_1, \cdots, v_n$ with corresponding real eigenvalues $\lambda_1, \cdots, \lambda_n$ (not necessarily distinct). Find $x(t)$ in terms of the eigenvalues and eigenvectors of A, and give a geometric interpretation of the result when $x_0 = v_i$ for some $i = 1, \cdots, n$.

Solution :

Since $\{v_1, \cdots, v_n\}$ is linearly independent, it must be a basis for $\mathbb{R}^n$. Thus, there exist scalars $a_1, \cdots, a_n$ such that:

$$x_0 = a_1v_1 + \cdots + a_nv_n$$

Substituting into (3.8), we have:

$$\begin{aligned} x(t) &= e^{tA}x_0 \\ &= e^{tA}(a_1v_1 + \cdots + a_nv_n) \\ &= a_1e^{\lambda_1 t}v_1 + \cdots + a_ne^{t\lambda_n t}v_n \end{aligned}$$

In particular, $x_0 = v_i$ corresponds to the case that $a_j = \delta_{ij}$, where the Kronecker delta δ_{ij} equals 1 if $i = j$ and equals 0 otherwise. In this case, the above equation becomes:

$$x(t) = v_ie^{\lambda_i t}$$

Thus, the trajectory diverges if $\Re\lambda_i > 0$, and converges otherwise.

Example (**Discussion 7, Problem 5**). Suppose $A \in \mathbb{R}^{2\times 2}$ has eigenvalues $\lambda, \overline{\lambda} \in \mathbb{C}$, where $\overline{\lambda}$ denotes the complex conjugate of λ. Let $v \in \mathbb{C}^n$ be the (complex) eigenvector of A corresponding to the eigenvalue λ, and suppose:

$$\begin{aligned} \lambda &= \sigma + j\omega \\ v &= v_1 + jv_2, \end{aligned}$$

where $\sigma, \omega \in \mathbb{R}$ and $v_1, v_2 \in \mathbb{R}^n$.

Now, consider:

$$\dot{x} = Ax, \qquad x(0) = \frac{1}{2}(v + \overline{v}),$$

where $\overline{v}$ denotes the (term-wise) complex conjugate of v. Find an expression for the trajectory $x(t)$ in terms of $v_1, v_2, \sigma, \omega, t$, and give a geometric interpretation of $x(t)$.

Solution :

Using (3.8), we have:

$$\begin{aligned} x(t) = e^{tA}x_0 = e^{tA} \cdot \frac{1}{2}(v + \overline{v}) &= \frac{1}{2}e^{tA}v + \frac{1}{2}e^{tA}\overline{v} \\ &= \frac{1}{2}e^{t\lambda}v + \frac{1}{2}e^{t\overline{\lambda}}\overline{v} \\ &= \frac{1}{2}e^{(\sigma+j\omega)t}(v_1 + jv_2) + \frac{1}{2}e^{(\sigma-j\omega)t}(v_1 - jv_2) \\ &= \frac{1}{2}e^{\sigma t}(e^{j\omega t}v_1 + e^{-j\omega t}v_1 + je^{j\omega t}v_2 - je^{-j\omega t}v_2 \\ &= e^{\sigma t}(\cos(\omega t)v_1 + \sin(\omega t)v_2) \end{aligned}$$

Example (**Discussion 7, Problem 10**). Given:

$$A = \begin{bmatrix} -3 & 1 \\ 0 & -2 \end{bmatrix}$$

Answer the following questions.

1. Find the characteristic polynomial of A.
2. Express A^4 as the lowest order polynomial in A.
3. Find e^{tA} by the Cayley-Hamilton Theorem; that is, show that $e^{tA} = a_0(t)I_2 + a_1(t)A$.

Solution :

1. We have:
$$\chi_A(\lambda) = (\lambda + 3)(\lambda + 2)$$
2. The Cayley-Hamilton Theorem implies that:
$$O = \chi_A(A) = (A + 3I_2)(A + 2I_2) = A^2 + 5A + 6I_2$$
Using long division, we have:
$$\begin{aligned} A^4 &= (A^2 + 5A + 6I_2)(A^2 - 5A + 19I_2) + (-65A - 114I_2) \\ &= -65A - 114I_2 \end{aligned}$$
3. Note that there exist t-dependent coefficients $a_0(t), a_1(t)$ such that:
$$e^{t\lambda} = q(\lambda) \cdot (\lambda + 3)(\lambda + 2) + a_1(t)\lambda + a_0(t)$$

Setting $\lambda = -3$ and $\lambda = -2$, we have:

$$\begin{aligned} e^{-3t} &= 3a_1(t) + a_0(t) \\ e^{-2t} &= 2a_1(t) + a_0(t), \end{aligned}$$

which yields $a_1(t) = e^{-2t} - e^{-3t}$ and $a_2(t) = 3e^{-2t} - 2e^{-3t}$.

Substituting λ with A, we have:

$$\begin{aligned} e^{tA} &= a_1(t)A + a_0 I_2 \\ &= (e^{-2t} - e^{-3t}) \begin{bmatrix} s-3 & 1 \\ 0 & -2 \end{bmatrix} + (3e^{-2t} - 2e^{-3t}) \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} e^{-3t} & e^{-3t} - e^{-2t} \\ 0 & e^{-2t} \end{bmatrix} \end{aligned}$$

4.3 Lecture 13

If a given square matrix $A \in \mathbb{R}^{n\times n}$ is diagonalizable, then each of its eigenvectors $\{v_i, i = 1, \cdots, n\}$ will remain *invariant* with respect to their original one-dimensional subspaces, i.e.:

$$Av_i \in \text{span}\{v_i\}$$

for each $i = 1, \cdots, k$. However, this does not hold in general for non-diagonalizable square matrices. This motivates the following definition.

Definition 4.5 (**Subspace Invariance**). *Given a vector space V and a linear operator $A : V \to V$, a subsspace $M \leq V$ is **invariant under** A, or A-**invariant** if, for each $x \in M$, we have $Ax \in M$.*

It can be easily checked that the null space and range space of any linear operator A are A-invariant. This is detailed in the lemma below. The proofs are left as exercises to the reader.

Lemma 4.6. *Let A be a linear operator. If $\lambda \in \sigma(A)$, and $p(s)$ is any polynomial, the following subspaces are A-invariant:*

1. $N(A)$
2. $R(A)$
3. $N(A - \lambda I)$
4. $N(p(A))$

Theorem 4.7. *If $M_1, M_2 \in V$ are A-invariant, so are $M_1 + M_2$ and $M_1 \cup M_2$.*

Definition 4.8 (**Direct Sum**). *Let V be a vector space, and let $M_1, \cdots, M_k \leq V$. Then V is said to be the **direct sum** of $M_1, \cdots, M_k$, if, for each $i = 1, \cdots, k$, denoted by:*

$$V = M_1 \oplus M_2 \oplus \cdots \oplus M_k,$$

if, for each $x \in V$, there exists a unique $x_i \in M_i$ for each $i = 1, \cdots, k$ such that:

$$x = x_1 + \cdots + x_k$$

Theorem 4.9 (**Equivalent Definition for Direct Sum of Subspaces**). *Let V be a vector space, and let $M_1, \cdots, M_k \leq V$. Then $V = M_1 \oplus \cdots \oplus M_k$ if and only if:*

1. $V = M_1 + \cdots + M_k$, *and*
2. $M_1 \bigcap \cdots \bigcap M_k = \{0\}$.

The above definitions for subspace invariance and direct sum to make the following observation. Let V be a vector space, and let A be a linear operator on V. If V is the direct sum of a sequence of A-invariant subspaces $M_1, \cdots, M_k$, then there exists a choice of basis $\mathcal{B}$ for V, constructed by taking the union of bases for each M_i, such that the matrix representation of A with respect to $\mathcal{B}$ is block diagonal.

For example, for the $k = 2$ case, if $\{b_1, \cdots, b_k\}$ is a basis for M_1, and $\{b_{k+1}, \cdots, b_n\}$ is a basis for M_2, then:

$$\{b_1, \cdots, b_k, b_{k+1}, \cdots, b_n\}$$

A version of the above result is discussed below.

Theorem 4.10 (**2nd Representation Theorem**). *Let $V = M_1 \oplus M_2$ be a finite-dimensional vector space. If M_1 is A-invariant, then there exists a basis $\mathcal{B}$ of V such that the matrix representation of A with respect to $\mathcal{B}$ has the form:*

$$[A]_{\mathcal{B}} = \begin{bmatrix} A_{11} & A_{12} \\ O & A_{22} \end{bmatrix} \in \mathbb{F}^{n \times n},$$

where $A_11 \in \mathbb{F}^{k \times k}$, $A_{12} \in \mathbb{F}^{k \times (n-k)}$, and $A_{22} \in \mathbb{F}^{(n-k) \times (n-k)}$. In other words, the first k columns of A have zeros in the last $(n-k)$ rows.

Proof. Essentially, we wish to show that $\left([A]_{\mathcal{B}}\right)_{ji} = 0$ for each $i = 1, \cdots, k$ and $j = k+1, \cdots, n$.

Let $\{b_1, \cdots, b_k\}$ and $\{b_{k+1}, \cdots, b_n\}$ be bases for M_1 and M_2, respectively. Then $\mathcal{B} \equiv \{b_1, \cdots, b_n\}$ forms a basis for $\mathcal{V}$. Then, for each $b_i, i = 1, \cdots, k$, we have the following equality from the definition of matrix representations:

$$Ab_i = \sum_{j=1}^{n} \left([A]_{\mathcal{B}}\right)_{ji} b_j$$

where e_i denotes the i-th standard vector in $\mathbb{F}^n$. Since M_1 is A-invariant, we have $Ab_i \in A$, and since vector representations with respect to a given basis are unique, we must have:

$$Ab_i = \sum_{j=1}^{k} \left([A]_{\mathcal{B}}\right)_{ji} b_j$$

This implies that, for each $i = 1, \cdots, k$:

$$\sum_{j=k+1}^{n} \left([A]_{\mathcal{B}}\right)_{ji} b_j = 0$$

Since $\{b_j, j = k+1, \cdots, n\}$ forms a basis for M_2, we have $\left([A]_{\mathcal{B}}\right)_{ji} = 0$ for each $i = 1, \cdots, k$ and $j = k+1, \cdots, n$. This completes the proof. ■

If M_1, M_2 were both A-invariant, then the above argument can be repeated to show that $A_{12} = 0$, so:

$$[A]_{\mathcal{B}} = \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix} \in \mathbb{F}^{n\times n},$$

where $A_{11} \in \mathbb{F}^{k\times k}$ and $A_{22} \in \mathbb{F}^{(n-k)\times(n-k)}$, as before. This is simply the observation described above.

Corollary 4.11. *Given a diagonalizable matrix $A \in \mathbb{R}^{n\times n}$ with n distinct eigenvalues $\lambda_1, \cdots, \lambda_n$:*

$$\mathbb{C}^n = N(A - \lambda_1 I) \oplus N(A - \lambda_2 I) \oplus \cdots \oplus N(A - \lambda_n I)$$

In general, suppose $A \in \mathbb{R}^{n\times n}$, not necessarily diagonalizable, has a characteristic equation of the form:

$$\begin{aligned} \chi_A(s) &= (s - \lambda_1)^{d_1}(s - \lambda_2)^{d_2} \cdot \cdots \cdot (s - \lambda_\sigma)^{d_1} \\ &= \prod_{i=1}^{\sigma} (s - \lambda_i)^{d_i} \end{aligned}$$

where d_i is the algebraic multiplicity of λ_i, we have $d_1 + \cdots d_k = n$, and σ is the number of distinct eigenvalues of A (clearly, $\sigma \leq n$.

Definition 4.12 (**Annihilating Polynomial**)**.** *Given a linear operator A, an annihilating polynomial of A is a finite-dimensional polynomial $f(s)$ such that:*

$$f(A) = 0$$

Definition 4.13 (**Minimal Polynomial**)**.** *Given a linear operator A, a **minimal polynomial** $\psi_A(s)$ of A is a non-zero polynomial of the least degree, with leading coefficient 1, that annihilates A, i.e. it is a polynomial of the last degree such that:*

$$\psi_A(s) = O_{n\times n}$$

Below, we show that the minimal polynomial must be a factor of any annihilating factor. This can be sued to show that the minimal polynomial is unique.

Theorem 4.14. *Let A be a linear operator. If $p(A)$ is an annihilating polynomial of A, and $\psi_A(s)$ is a minimal polynomial of A, then $\psi_A(s)$ divides $p(A)$.*

Proof. Since $p_A(s)$ annihilates A, it must have higher (or the same) degree as $\chi_A(s)$, a polynomial of least degree that annihilates A. Thus, there exist matrix polynomials $q(s), r(s)$, with $\deg(r(s)) \leq \deg(\psi_A(s))$, such that:

$$\begin{aligned} p_A(s) &= \psi_A(s)\dot{q}(s) + r(s) \\ \Rightarrow p_A(A) &= \psi_A(A)\dot{q}(A) + r(A). \end{aligned}$$

Since $p_A(A) = \psi_A(A) = 0$, we have $r(A) = 0$. If $r(s)$ is non-zero, then it is a non-zero polynomial of degree less than $\deg\psi_A(s)$ that annihilates A, a contradiction. Thus, $r(s) = 0$, so $p_A(A) = \psi_A(A)\dot{q}(A)$. The proof is done. ■

Corollary 4.15. *The minimal polynomial $\psi_A(s)$ of any linear operator A is unique.*

Proof. Suppose $\psi_{A,1}(s)$ and $\psi_{A,2}(s)$ are both minimal polynomials of A. Then, by the above argument, there exist polynomials $q_1(s), q_2(s)$ such that:

$$\begin{aligned} \psi_{A,1}(s) &= q_1(s) \cdot \psi_{A,2}(s) \\ \psi_{A,2}(s) &= q_2(s) \cdot \psi_{A,1}(s) \\ \Rightarrow \psi_{A,1}(s) &= q_1(s) \cdot \psi_{A,2}(s) \\ &= q_1(s)q_2(s) \cdot \psi_{A,1}(s) \\ \Rightarrow q_1(s)q_2(s) &= 1 \end{aligned}$$

Since $\psi_{A,1}(s)$ and $\psi_{A,2}(s)$ both have leading coefficient 1, so must $q_1(s)$ and $q_2(s)$. The fact that $q_1(s)q_2(s) = 1$ implies that $q_1(s)$ and $q_2(s)$ are both constant. Combining these two facts, we have:

$$q_1(s) = q_2(s) = 1,$$

and so $\psi_{A,1}(s) = \psi_{A,2}(s)$. ■

The corollary below shows that $\chi_A(s)$ and $\psi_A(s)$ must have the same roots.

Corollary 4.16. *Let $A \in \mathbb{C}^{n\times n}$ be a linear operator with distinct eigenvalues $\lambda_1, \cdots, \lambda_\sigma$, where $\sigma \leq n$. Then the minimal polynomial $\chi_A(s)$ of A has the form:*

$$\psi_A(s) = (s-\lambda_1)^{m_1} \cdot (s-\lambda_2)^{m_2} \cdot \cdots \cdot (s-\lambda_\sigma)^{m_\sigma},$$

where $1 \leq m_i \leq a_i$ for each $i = 1, \cdots, \sigma$.

Proof. Taking $p(s) = \chi_A(s)$ (the characteristic polynomial of A) in the above theorem, we find that $\chi_A(s)$ is a factor of $p(s)$. Since $\chi_A(s)$ has the form:

$$\chi_A(s) = (s-\lambda_1)^{a_1} \cdot \cdots \cdot (s-\lambda_\sigma)^{a_\sigma}$$

the minimal polynomial $p(s)$ must have the form:

$$\chi_A(s) = (s-\lambda_1)^{m_1} \cdot \cdots \cdot (s-\lambda_\sigma)^{m_\sigma}$$

where $0 \leq m_i \leq a_i$ for each $i = 1, \cdots, \sigma$.

It remains to show that $m_i \geq 1$ for each $i = 1, \cdots, \sigma$, i.e. that each λ_i is a root of $\psi_A(s)$. Let v_i be an eigenvector of A with corresponding eigenvalue λ_i. Then, since $\psi_A(A) = 0$, we have:

$$0 = \psi_A(A)v_i = \psi_A(\lambda_i)v_i,$$

where the replacement of A with λ_i in the second equality follows from the observation that $A^k v_i = \lambda_i^k v_i$ for each $k \in \mathbb{N}$. Since $v_i \neq 0$, we have $\psi_A(\lambda_i) = 0$, which shows that λ_i is a root of $\psi_A(s)$. The proof is done. ■

Theorem 4.17. *Given any matrix $A \in \mathbb{R}^{n\times n}$ with $\sigma \leq n$ distinct eigenvalues $\lambda_1, \cdots, \lambda_\sigma$:*

$$\mathbb{C}^n = N(A-\lambda_1 I)^{m_1} \oplus N(A-\lambda_2 I)^{m_2} \oplus \cdots \oplus N(A-\lambda_n I)^{m_n}$$

Proof. Fix $x \in \mathbb{C}^n$, and observe that:

$$\begin{aligned}
&\psi_A(s) = (s-\lambda_1)^{m_1} \cdot (s-\lambda_2)^{m_2} \cdot \cdots \cdot (s-\lambda_\sigma)^{m_\sigma} \\
\Rightarrow &\frac{1}{\psi_A(s)} = \frac{1}{(s-\lambda_1)^{m_1} \cdot (s-\lambda_2)^{m_2} \cdot \cdots \cdot (s-\lambda_\sigma)^{m_\sigma}} \\
&\qquad = \frac{n_1(s)}{(s-\lambda_1)^{m_1}} + \frac{n_2(s)}{(s-\lambda_2)^{m_2}} + \cdots + \frac{n_\sigma(s)}{(s-\lambda_\sigma)^{m_\sigma}} \\
\Rightarrow &I = \frac{1}{\psi_A(A)} \cdot \psi_A(A) \\
&\qquad = n_1(A) \cdot \frac{\psi_A(A)}{(A-\lambda_1 I)^{m_1}} + \cdots + n_\sigma(A) \cdot \frac{\psi_A(A)}{(A-\lambda_\sigma I)^{m_\sigma}} \\
\Rightarrow &x = n_1(A)\frac{\psi_A(A)}{(A-\lambda_1 I)^{m_1}}x + \cdots + n_\sigma(A)\frac{\psi_A(A)}{(A-\lambda_\sigma I)^{m_\sigma}}x
\end{aligned}$$

where $n_1(s), \cdots, n_\sigma(s)$ are functions of s with degree less than $m_1, \cdots, m_\sigma$, respectively. For each $i = 1, \cdots, n$:

$$(A-\lambda_i I)^{m_i}\left(n_i(A)\frac{\psi_A(A)}{(A-\lambda_i I)^{m_\sigma}}x\right) = n_i(A)\psi_A(A)x = 0,$$

since $\psi_A(A) = 0$. Thus, x is written as a linear combination of elements in $N(A-\lambda_i I)^{m_i}$, i.e.:

$$\mathbb{C}^n = N(A-\lambda_1 I)^{m_1} + N(A-\lambda_2 I)^{m_2} + \cdots + N(A-\lambda_n I)^{m_n}$$

To complete the proof, it remains to show that $N(A-\lambda_i I) \bigcap N(A-\lambda_j I) = \{0\}$ for each $i, j \in \{1, \cdots, \sigma\}$. Fix one such pair of i, j, and suppose without loss of generality that $1 \leq m_i \leq m_j$. Let $v \in N(A-\lambda_i)^{m_i}$, and define:

$$\overline{m_i} = \min\{m | v \in N(A-\lambda_i)^m\}$$

The definition of m shows that $\{m | v \in N(A-\lambda_i)^m\}$ is bounded above by m_i, so $\overline{m_i}$ exists. Now:

$$\begin{aligned}
0 = (A-\lambda_i I)^{\overline{m_i}}v &= (A-\lambda_i I)(A-\lambda_i I)^{\overline{m_i}-1}v \\
\Rightarrow A(A-\lambda_i I)^{\overline{m_i}-1}v &= \lambda_i(A-\lambda_i I)^{\overline{m_i}-1}v
\end{aligned}$$

In other words, $(A-\lambda_i I)^{m_i-1}$ is an eigenvector of A with corresponding eigenvalue λ_i (By definition of $\overline{m}_i$, we have $(A-\lambda_i I)^{m_i-1} \neq 0$). We thus have:

$$\begin{aligned}
(A-\lambda_j I)^{m_2}v &= (A-\lambda_j I)^{m_2-\overline{m_1}+1}(A-\lambda_j I)^{\overline{m_1}-1}v \\
&= (\lambda_i-\lambda_j)^{m_2-\overline{m_1}+1}(A-\lambda_j I)^{\overline{m_1}-1}v \neq 0
\end{aligned}$$

since $(\lambda_i-\lambda_j)^{m_2-\overline{m_1}+1} \neq 0$ and $(A-\lambda_j I)^{\overline{m_1}-1}v \neq 0$.

■

Definition 4.18 (**Algebraic and Geometric Multiplicities, Grade**). *Let $A \in \mathbb{R}^{n\times n}$ be given, and suppose A has characteristic polynomial given by:*

$$\chi_A(\lambda) = (\lambda - \lambda_1)^{d_1} \cdot \cdots \cdot (\lambda - \lambda_\sigma)^{a_\sigma}$$

where $\lambda_1, \cdots, \lambda_\sigma$ are the distinct eigenvalues of A, i.e. $\sigma(A) = \{\lambda_1, \cdots, \lambda_\sigma\}$ and minimal polynomial given by:

$$\psi_A(\lambda) = (\lambda - \lambda_1)^{m_1} \cdot \cdots \cdot (\lambda - \lambda_\sigma)^{m_\sigma}$$

Then, we have the following definitions:

1. **Algebraic Multiplicity:**

   *d_i is called the **algebraic multiplicity** of λ_i. Note that $\sum_{i=1}^{\sigma} d_i = 1$*

2. **Geometric Multiplicity:**

   *$q_i \equiv \dim(N(A - \lambda_i))$ is called the **geometric multiplicity** of λ_i, and describes the maximum number of linearly independent eigenvectors corresponding to the eigenvalue λ_i.*

3. **Grade:**

   *$m_i \equiv \dim(N(A - \lambda_i))$ is called the **grade** of λ_i. It describes the largest Jordan block of A corresponding to λ_i, since it takes m_i multiplications of $(A - \lambda_i)$ to annihilate the blocks.*

Example. Consider the algebraic multiplicities, geometric multiplicities, and grades of the following square matrices (observe that they are already in Jordan form), with $\lambda \in \mathbb{C}$ arbitrarily given:

$$A_1 = \begin{bmatrix} \lambda & 1 & 0 & 0 \\ 0 & \lambda & 0 & 0 \\ 0 & 0 & \lambda & 1 \\ 0 & 0 & 0 & \lambda \end{bmatrix}, \qquad A_2 = \begin{bmatrix} \lambda & 1 & 0 & 0 \\ 0 & \lambda & 1 & 0 \\ 0 & 0 & \lambda & 0 \\ 0 & 0 & 0 & \lambda \end{bmatrix}$$

By observation, we can compile the following table. The algebraic multiplicity of λ is 4 for both A_1 and A_2, since they are both 4×4 matrices with whose only eigenvalue is λ. The geometric multiplicity of λ is 2 for both A_1 and A_2 since they both have two Jordan blocks. Finally, the grades of λ for A_1 and A_2 are 2 and 3, respectively, since the maximum size of the Jordan blocks corresponding to λ in A_1 and A_2 are 2×2 and 3×3, respectively.

Parameter	A_1	A_2
Algebraic Multiplicity	4	4
Geometric Multiplicity	2	2
Grade of λ	2	3

Theorem 4.19 (**Spectral Mapping Theorem**). *Given a square matrix $A \in \mathbb{R}^{n\times n}$, and a polynomial $f(\lambda)$, we have:*

$$\sigma(f(A)) = f(\sigma(A)),$$

i.e. the spectrum of $f(A)$ is the set of $f(\lambda)$ for each eigenvalue λ of A.

Proof. Explicitly express f as:

$$f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$$

Let $\lambda \in \sigma(A)$. Then there exists some eigenvector v of A such that $Av = \lambda v$. Thus:

$$\begin{aligned} f(A)v &= (a_n A^n + a_{n-1}A^{n-1} + \cdots + a_1 A + a_0 I)v \\ &= (a_n \lambda^n + a_{n-1}\lambda^{n-1} + \cdots + a_1 \lambda + a_0)v \\ &= f(\lambda)v \end{aligned}$$

This establishes the theorem. ■

Remark. This result can be extended to all analytic functions (not just polynomials), via Taylor expansion.

Consider the following simple example of a 2×2 matrix. Suppose:

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

It can be verified that:

$$J^n = \begin{bmatrix} \lambda^n & n\lambda^{n-1} \\ 0 & \lambda^n \end{bmatrix}$$

Then, for the matrix exponential e^{tJ}, we have

$$\begin{aligned} e^{tJ} &= I + tJ + \frac{(tJ)^2}{2!} + \cdots + \frac{(tJ)^n}{n!} + \cdots \\ &= \begin{bmatrix} e^{\lambda t} & b \\ 0 & e^{\lambda t} \end{bmatrix}, \end{aligned}$$

where $b \in \mathbb{R}$ can be determined by using the formula above for J^n:

$$\begin{aligned} b &= t + 2\lambda\frac{t^2}{2!} + \cdots + n\lambda^{n-1} \cdot \frac{t^n}{n!} + \cdots \\ &= te^{\lambda t} \end{aligned}$$

Thus:

$$e^{tJ} = \begin{bmatrix} e^{\lambda t} & te^{\lambda t} \\ 0 & e^{\lambda t} \end{bmatrix}$$

In general, we can show that, if $J \in \mathbb{R}^{n \times n}$ is given by:

$$J = \begin{bmatrix} \lambda & 1 & 0 & \cdots & 0 & 0 \\ 0 & \lambda & 1 & \cdots & 0 & 0 \\ 0 & 0 & \lambda & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda & 1 \\ 0 & 0 & 0 & \cdots & 0 & \lambda \end{bmatrix}$$

Then, we have (by induction):

$$e^{tJ} = \begin{bmatrix} e^{t\lambda} & te^{t\lambda} & \frac{t^2}{2!}e^{t\lambda} & \cdots & \frac{t^{n-2}}{(n-2)!}f^{(n-2)}(\lambda) & \frac{t^{n-1}}{(n-1)!}f^{(n-1)}(\lambda) \\ 0 & e^{t\lambda} & te^{t\lambda} & \cdots & \frac{t^{n-3}}{(n-3)!}f^{(n-3)}(\lambda) & \frac{t^{n-2}}{(n-2)!}f^{(n-2)}(\lambda) \\ 0 & 0 & e^{t\lambda} & \cdots & \frac{t^{n-4}}{(n-4)!}f^{(n-4)}(\lambda) & \frac{t^{n-3}}{(n-3)!}f^{(n-3)}(\lambda) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & e^{t\lambda} & te^{t\lambda} \\ 0 & 0 & 0 & \cdots & 0 & e^{t\lambda} \end{bmatrix}$$

The above generalized result can be proven by the following theorems.

Lemma 4.20. *Let $A \in \mathbb{R}^{n\times n}$ be a square matrix, with minimal polynomial:*

$$\psi_A(s) = (s-\lambda_1)^{m_1} \cdot \cdots \cdot (s-\lambda_\sigma)^{m_\sigma}$$

Let $n = \sum_{i=1}^{\sigma} d_i = n$, and let $h(s)$ be an $(n-1)$-th order polynomial. Then $f(A) = h(A)$ if and only if:

$$f^{(k)}(\lambda_i) = h^{(k)}(\lambda_i), \qquad \forall\, k = 0, 1, \cdots, m_i - 1, \quad i = 1, \cdots \sigma,$$

where $f^{(k)}(\lambda_i)$ and $h^{(k)}(\lambda_i)$ respectively denote the k-th derivative of f and g, evaluated at λ_i.

Proof. The proof follows by Taylor expansion at each $\lambda_i \in \sigma(A)$. ■

Theorem 4.21 (**Functions of a Matrix**). *Consider an $n \times n$ Jordan block given by:*

$$J = \begin{bmatrix} \lambda & 1 & 0 & \cdots & 0 & 0 \\ 0 & \lambda & 1 & \cdots & 0 & 0 \\ 0 & 0 & \lambda & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda & 1 \\ 0 & 0 & 0 & \cdots & 0 & \lambda \end{bmatrix}$$

Then:

$$f(J) = \begin{bmatrix} f(\lambda) & f'(\lambda) & \frac{1}{2!}f''(\lambda) & \cdots & \frac{1}{(n-2)!}f^{(n-2)}(\lambda) & \frac{1}{(n-1)!}f^{(n-1)}(\lambda) \\ 0 & f(\lambda) & f'(\lambda) & \cdots & \frac{1}{(n-3)!}f^{(n-3)}(\lambda) & \frac{1}{(n-2)!}f^{(n-2)}(\lambda) \\ 0 & 0 & f(\lambda) & \cdots & \frac{1}{(n-4)!}f^{(n-4)}(\lambda) & \frac{1}{(n-3)!}f^{(n-3)}(\lambda) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & f(\lambda) & f'(\lambda) \\ 0 & 0 & 0 & \cdots & 0 & f(\lambda) \end{bmatrix}$$

Proof. From the above lemma, we have:

$$f(J) = \sum_{k=0}^{n-1} f^{(k)}(\lambda) \cdot \frac{1}{k!}(x-\lambda)^k$$

Thus, by substituting x with J, we find:

$$f(J) = \sum_{k=0}^{\infty} f^{(k)}(\lambda) \cdot \frac{1}{k!}(J - \lambda I)^k$$

The proof is completed by observing that, for each $m = 1, \cdots, n$, we have:

$$[(J - \lambda I)^m]_{ij} = \delta_{i,j-m}$$

■

Note that the Jordan form $J = P^{-1}AP$ is unique only up to permutations of the Jordan blocks. This can be seen by interchanging the order in which different Jordan chains (i.e. sets containing a single eigenvector and generalized eigenvectors associated with that eigenvector) appear in P.

Example (**From an Old Prelim**).

1. A matrix A has minimal polynomial $(s - \lambda_1)^2(s - \lambda_2)^3$. Find $\cos(e^A)$ as a polynomial in A.

2. Now, assume further that A has characteristic polynomial $(s - \lambda_1)^2(s - \lambda_2)^3$, and that it has four linearly independent eigenvectors. Write down the Jordan form J of this matrix, and write down $\cos(e^A)$ explicitly.

Solution:

1. Since the minimal polynomial, by definition, is the (unique) polynomial of the lowest degree with leading coefficient 1 that annihilates A. To that end, let a (infinite) polynomial $q(\lambda)$ and constants c_4, c_3, c_2, c_1, c_0 be given such that:

$$\cos(e^\lambda) = q(\lambda) \cdot (\lambda - \lambda_1)^2(\lambda - \lambda_2)^3 + c_4\lambda^4 + c_3\lambda^3 + c_2\lambda^2 + c_1\lambda + c_0$$

The five constants can be found by differentiating $\cos(e^\lambda)$ with respect to λ twice times, and substituting $\lambda = \lambda_1$ to $\cos(e^\lambda)$ and its first derivative, and substituting $\lambda = \lambda_2$ to $\cos(e^\lambda)$ and its first and second derivatives. The derivatives of $\cos(e^\lambda)$ are:

$$\frac{d}{d\lambda}\cos(e^\lambda) = -e^\lambda \sin(e^\lambda)$$
$$\frac{d^2}{d\lambda^2}\cos(e^\lambda) = -e^\lambda \sin(e^\lambda) - e^{2\lambda}\cos(e^\lambda)$$

So, substituting $\lambda = \lambda_1$ and $\lambda = \lambda_2$ in the relevant equations, we have:

$$\begin{aligned}
\cos(e^{\lambda_1}) &= c_4\lambda_1^4 + c_3\lambda_1^3 + c_2\lambda_1^2 + c_1\lambda_1 + c_0 \\
\cos(e^{\lambda_2}) &= c_4\lambda_2^4 + c_3\lambda_2^3 + c_2\lambda_2^2 + c_1\lambda_2 + c_0 \\
-e^{\lambda_1}\sin(e^{\lambda_1}) &= 4c_4\lambda_1^3 + 3c_3\lambda_1^2 + 2c_2\lambda_1 + c_1 \\
-e^{\lambda_2}\sin(e^{\lambda_2}) &= 4c_4\lambda_2^3 + 3c_3\lambda_2^2 + 2c_2\lambda_2 + c_1 \\
-e^{\lambda_2}\sin(e^{\lambda_2}) - e^{2\lambda_2}\cos(e^{\lambda_2}) &= 12c_4\lambda_2^2 + 6c_3\lambda_2 + 2c_2
\end{aligned}$$

Thus, c_4, c_3, c_2, c_1, c_0 can be solved as functions of t. We thus have:

$$A = c_4A^4 + c_3A^3 + c_2A^2 + c_1A + c_0I$$

2. Note that the exponent of each term in minimal polynomials gives the size of the largest Jordan block corresponding to each eigenvalue; in this case, $m_1 = 2$ and $m_2 = 3$. Since each linearly independent eigenvector of A gives rise to one Jordan block, the total number of Jordan blocks in the Jordan form of A is 4 (In other words, they occupy a 4×4 matrix in the Jordan form of A). On the other hand, the algebraic multiplicities corresponding to each eigenvalue of A is the total size of all Jordan blocks corresponding to that eigenvalue.

   In the context of this problem, the above information implies that:

   - For λ_1, the largest Jordan block is of size 2, and the sum of the sizes of the Jordan blocks is 5.
   - For λ_1, the largest Jordan block is of size 3, and the sum of the sizes of the Jordan blocks is 3.
   - In the Jordan form of A, the eigenvalues λ_1 and λ_2 have a total of four Jordan blocks.

   The second fact in the above list indicates that there exists only one (3×3) Jordan block corresponding to λ_2. Thus, there exist three Jordan blocks corresponding to λ_1. These blocks have maximum size 2 and total size 5, so there must be two blocks of size 2 and one block of size 1. In short, the Jordan form of A looks as follows:

$$J = \left[\begin{array}{c|cc|cc|ccc} \lambda_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & \lambda_1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \lambda_1 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & \lambda_1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \lambda_1 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & \lambda_2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \lambda_2 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_2 \end{array}\right]$$

   By Theorem 4.21, for any analytic function $f(\cdot)$, we have:

$$f(J) = \left[\begin{array}{c|cc|cc|ccc} f(\lambda_1) & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & f(\lambda_1) & f'(\lambda_1) & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & f(\lambda_1) & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & f(\lambda_1) & f'(\lambda_1) & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & f(\lambda_1) & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & f(\lambda_2) & f'(\lambda_2) & \frac{1}{2}f''(\lambda_2) \\ 0 & 0 & 0 & 0 & 0 & 0 & f(\lambda_2) & f'(\lambda_2) \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & f(\lambda_2) \end{array}\right]$$

   The solution is completed by substituting $f(x) = \cos(e^x)$ and its first and second derivatives into the expression above. The resulting matrix is a similarity transform away from

$f(A)$, with respect to the same invertible matrix P (with columns/rows consisting of the column/row eigenvectors of A) that connects A to J via the similarity transformation:

$$A = PJP^{-1}$$

More examples are furnished in the Discussion questions listed below.

Corollary 4.22. *For any $A \in \mathbb{R}^{n\times n}$, the characteristic and minimal polynomials of A and A^T are the same.*

Proof. First, since the determinant is invariant with respect to taking the transpose of the matrix:

$$\chi_A(\lambda) = \det(A - \lambda I) = \det(A^T - \lambda I) = \chi_{A^T}(\lambda)$$

Next, suppose the minimal polynomial of A is of the form:

$$m_A(\lambda) = (\lambda - \lambda_1)^{m_1} \cdot \cdots \cdot (\lambda - \lambda_\sigma)^{m_\sigma} = 0,$$

The Cayley-Hamilton Theorem implies that $m_A(A) = O_n$, so:

$$\begin{aligned} O_n &= \big(m_A(A)\big)^T \\ &= \big((A - \lambda_1 I_n)^{m_1} \cdot \cdots \cdot (A - \lambda_\sigma I_n)^{m_\sigma}\big)^T \\ &= (A^T - \lambda_\sigma I_n)^{m_\sigma} \cdot \cdots \cdot (A^T - \lambda_1 I_n)^{m_1} \\ &= (A^T - \lambda_1 I_n)^{m_1} \cdot \cdots \cdot (A^T - \lambda_\sigma I_n)^{m_\sigma} \\ &= m_A(A^T) \end{aligned}$$

It follows that $m_A(\lambda)$ annihilates A^T. But by definition of minimal polynomial, $m_{A^T}(\lambda)$ is the polynomial of least degree that annihilates A^T. Thus:

$$\deg(m_A(\lambda)) \leq m_{A^T}(\lambda)$$

By symmetry (or, by replacing A with A^T), we have $\deg(m_A(\lambda)) \geq m_{A^T}(\lambda)$. We conclude that $\deg(m_A(\lambda)) = m_{A^T}(\lambda)$. Since, by definition of minimal polynomial, both $m_A(\lambda)$ and $m_{A^T}(\lambda)$ have leading coefficient 1, they must be equal:

$$m_{A^T}(\lambda) = m_A(\lambda)$$

It follows that A and A^T have identical Jordan forms, up to a permutation of the Jordan blocks. ■

4.4 Lecture 13 Discussion

Example (**Discussion 8, Problems 1, 8, 10**). Consider the matrix A given below:

$$A = \begin{bmatrix} 3 & -2 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 2 \end{bmatrix}$$

Consider the following subspaces of $\mathbb{R}^3$:

$$M_1 = \text{span}\left\{ \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \right\}, \qquad M_2 = \text{span}\left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right\}$$

Answer the following questions:

1. Determine whether or not M_1 and M_2 are A-invariant.
2. Using M_1 and M_2, find a basis representation for A based on the 2nd Representation Theorem, i.e. with an block upper triangular form.

Solution:

1. Since $[0, 1, 0]^T \in M_1$, but:

$$A \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 3 & -2 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 2 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} -2 \\ 1 \\ -1 \end{bmatrix} \notin M_1,$$

   we conclude that M_1 is not an A-invariant subspace of $\mathbb{R}^3$.

   On the other hand, any vector in M_2 must have the form $(a, 0, b)^T$, and:

$$A \begin{bmatrix} a \\ 0 \\ b \end{bmatrix} = \begin{bmatrix} 3 & -2 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 2 \end{bmatrix} \begin{bmatrix} a \\ 0 \\ b \end{bmatrix} = \begin{bmatrix} 3a \\ 0 \\ 2b \end{bmatrix} \in M_2$$

   we conclude that M_1 is not an A-invariant subspace of $\mathbb{R}^3$.

2. Technically, A is already in block upper triangular form:

$$\left[\begin{array}{c|cc} 3 & -2 & 0 \\ \hline 0 & 1 & 0 \\ 0 & -1 & 2 \end{array}\right]$$

   This is due to the fact that, not only is M_2 an A-invariant subspace, so are the two subspaces spanned by the two linearly independent vectors used in the definition of M_2, namely $(1, 0, 0)^T$ and $(0, 0, 1)^T$. In particular, the block upper triangular form of A results

from the A-invariance of $(1,0,0)^T$. However, an (actual, not just block) upper triangular matrix representation of A can be derived by shuffling the vectors that generate M_1 and M_2 such that the two A-invariant vectors are placed first:

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

Example (**Discussion 8, Problems 4, 5**).

1. Find the characteristic polynomial and minimal polynomial of the following matrix:

$$A = \begin{bmatrix} 3 & 0 & 0 \\ 1 & 3 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

2. Verify that:

$$\dim\big(N(A - \lambda_i I)^{m_i}\big) = d_i$$

Solution:

1. Since A^T is in Jordan form, with a largest Jordan block of size 2, we have:

$$\chi_A(\lambda) = \chi_{A^T}(\lambda) = (\lambda - 3)^3$$
$$\psi_A(\lambda) = \psi_{A^T}(\lambda) = (\lambda - 3)^2$$

2. With regard to the left-hand side of the given equality, since $\lambda_i = 3$ and $m_i = 2$, we have, by the Cayley-Hamilton Theorem:

$$(A - \lambda_i I)^{m_i} = (A - 3I_3)^2 = O_3$$
$$\Rightarrow \dim\big(N(A - \lambda_i I)\big) = \dim\big(N(O^3)\big) = \dim(\mathbb{R}^3) = 3$$

Example (**Discussion 8, Problem 6**). A square matrix A has the following characteristic and minimal polynomials:

$$\chi_A(\lambda) = (\lambda - 1)^4(\lambda - 2)^2$$
$$\psi_A(\lambda) = (\lambda - 1)^2(\lambda - 2)$$

Answer the following questions:

1. What is the size of A?
2. Find the possible Jordan forms J of A, up to a permutation of the Jordan blocks.
3. Repeat the above question, this time with the additional constraint that A has 5 linearly independent eigenvectors.

Solution :

1. A square matrix A has dimension $n \times n$ if and only if its characteristic polynomial $\chi_A(\lambda)$ is of degree n. Here, $\deg(\chi_A(\lambda)) = 6$, so $A \in \mathbb{R}^{6\times 6}$.

2. From $\chi_A(\lambda)$ and $\psi_A(\lambda)$, we have:

$$\begin{aligned} d_1 &= 4, & d_2 &= 2 \\ m_1 &= 2, & m_2 &= 1 \end{aligned}$$

   It follows that the Jordan blocks corresponding to $\lambda = 1$ have maximum size 2 and total size 4, while the Jordan blocks corresponding to $\lambda = 2$ have maximum size 1 and total size 2. Thus, for λ_1, there could be two Jordan blocks of size 2, or one Jordan block of size 2 and two Jordan blocks of size 1. However, for λ_2, there can only be two Jordan blocks of size 1. The possibilities are as follows: (with the Jordan blocks corresponding to each eigenvalue placed in decreasing order of size along the diagnol):

$$J = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 \end{bmatrix}$$

$$\text{or} \qquad J = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 \end{bmatrix}$$

3. The number of linearly independent eigenvectors corresponding to each eigenvalue corresponds to the total number of Jordan blocks associated with that eigenvalue, as each eigenvector generates exactly one Jordan block. Thus, that A has 5 linearly independent eigenvectors corresponds to the presence of 5 Jordan blocks in J. This corresponds to the second matrix J written above.

4.5 Lecture 14

Below, we discuss input-output stability. Recall that, given a dynamical system described by equations of the form (3.2) and (3.3), we have:

$$y(t) = \int_0^t C(t)\Phi(t,\tau)B(\tau)u(\tau)\,d\tau + D(t)u(t)$$

Thus, in general, the relationship between the input and output of a system can be elucidated by rewriting the $y(t)$ and $u(t)$ as follows:

$$y(t) = \int_{-\infty}^t H(t,\tau)\,u(\tau)\,d\tau \tag{4.1}$$

Definition 4.23 (**Norms**). *In the ensuing discussion, the following norms will be used:*

1. $\|x\|_\infty = \max_i |x_i|$.
2. $\|A\|_{i,\infty} = \max_{1\le i\le m} \sum_{j=1}^n |a_{ij}|$
3. $\|u(\cdot)\|_\infty = \sup_{t\in\mathbb{R}} \|u(t)\|_\infty = \sup_{t\in\mathbb{R}}\{\max_{1\le j\le n_i} |u_j(t)|\}$
4. $\|y(\cdot)\|_\infty = \sup_{t\in\mathbb{R}} \|y(t)\|_\infty = \sup_{t\in\mathbb{R}}\{\max_{1\le j\le n_i} |y_j(t)|\}$
5. $L_\infty^{n_i} = \{u(\cdot) | \|u(\cdot)\|_\infty < \infty\}$
6. $L_\infty^{n_o} = \{y(\cdot) | \|y(\cdot)\|_\infty < \infty\}$

Note in particular that the matrix norm induced by the infinity norm is simply the max row sum.

Definition 4.24 (**Bounded Input, Bounded Output (BIBO) Stability**). *A system is* ***bounded-input, bounded-output (BIBO)*** *stable if there exists some $k \ge 0$ such that, for each $u(\cdot) \in L_\infty^{n_i}$, and each $t \ge 0$:*

$$\|y(\cdot)\|_\infty \le \|u(\cdot)\|_\infty \tag{4.2}$$

Remark.

1. (4.2) is BIBO stable if all bounded inputs produce bounded outputs. In fact, one can think of (4.2) as a linear operator $\mathcal{L} : L_\infty^{n_i} \to L_\infty^{n_0}$ such that:

$$\big(Lu(\cdot)\big)(t) \equiv= y(t) = \int_{-\infty}^t H(t,\tau)u(\tau)\,d\tau$$

2. Note that (4.2) specifies a linear relationship between y and u. Thus, as with induced norms, one can equivalently define a system to be BIBO stable if there exists some $k > 0$ such that, for each input $u(\cdot)$ with unit norm:

$$\|y(\cdot)\|_\infty \le k$$

3. A system is *not* BIBO stable if no $k > 0$ exists to satisfy (4.2), i.e. if there exists a sequence of unit-norm inputs $u^k(\cdot) \in L_\infty^{n_i}$, where $k = 1, 2, \cdots$, such that:

$$\|y^k(\cdot) \equiv (Lu^k(\cdot))\| > k$$

for each $k = 1, \cdots, n$.

4. On any finite-dimensional space, all norms are equivalent; thus, the inequality in the definition of BIBO stability can be given with respect to any norm without changing the definition.

Our main theorem involving BIBO stability will require the following lemma.

Lemma 4.25. *Given a matrix $H(t,\tau) = [h_{ij}(t,\tau)]_{n_0,n_i}$ dependent on two time instances, we have:*

$$\begin{aligned} &\sup_{t\in\mathbb{R}} \left\{ \int_{-\infty}^{t} \|H(t,\tau)\|_{i,\infty} d\tau \right\} < \infty \\ \iff &\sup_{t\in\mathbb{R}} \left\{ \int_{-\infty}^{t} |h_{ij}(t,\tau)|_{i,\infty} d\tau \right\} < \infty, \end{aligned}$$

for each $i = 1, \cdots, n_o$ and $j = 1, \cdots, n_i$.

Proof. The lemma follows straightforwardly from the fact that the induced matrix sup norm is in fact the maximum row sum, which implies:

$$|h_{ij}(t,\tau)| \leq \|H(t,\tau)\|_{i,\infty} \leq \sum_{i=1}^{n_0} \sum_{j=1}^{n_i} |h_{ij}(t,\tau)|.$$

■

Theorem 4.26. *The system given by (4.1) is BIBO stable if and only if:*

$$\sup_{t\in\mathbb{R}} \left\{ \int_{-\infty}^{t} \|H(t,\tau)\|_{i,\infty} d\tau \right\} < \infty$$

Equivalently, by the above lemma, (4.2) is BIBO stable if and only if:

$$\sup_{t\in\mathbb{R}} \left\{ \int_{-\infty}^{t} |h_{ij}(t,\tau)|_{i,\infty} d\tau \right\} < \infty$$

for each $i = 1, \cdots, n_0$ and $j = 1, \cdots, n_i$.

Proof.

" $\Rightarrow$ " Notice that:

$$\begin{aligned}
\|y(t)\|_\infty &= \left\| \int_{-\infty}^{t} H(t,\tau)u(\tau)\,d\tau \right\|_\infty \\
&\leq \int_{-\infty}^{t} \|H(t,\tau)u(\tau)\,\|_\infty\,d\tau \\
&\leq \int_{-\infty}^{t} \|H(t,\tau)\|_{i,\infty} \cdot \|u(\tau)\|_\infty\,d\tau \\
&= \int_{-\infty}^{t} \|H(t,\tau)\|_{i,\infty}\,d\tau \cdot \|u(\cdot)\|_\infty \\
&\leq \sup_{t\in\mathbb{R}} \left\{ \int_{-\infty}^{t} \|H(t,\tau)\|_{i,\infty} d\tau \right\} \cdot \|u(\cdot)\|_\infty
\end{aligned}$$

" $\Leftarrow$ " Suppose by contradiction that:

$$\sup_{t\in\mathbb{R}} \left\{ \int_{-\infty}^{t} \|H(t,\tau)\|_{i,\infty} d\tau \right\} = \infty$$

The above lemma implies that this is equivalent to assuming the existence of some $\alpha \in \{1,\cdots,n_o\}$ and $\beta \in \{1,\cdots,n_i\}$ such that:

$$\sup_{t\in\mathbb{R}} \int_{-\infty}^{t} ||h_{\alpha,\beta}(t,\tau)||d\tau$$

In other words, there exists a sequence of times $\{t_k\}_{k=1}^{\infty}$ such that:

$$\int_{-\infty}^{t_k} \|h_{\alpha,\beta}(t,\tau)\|_{i,\infty} d\tau > k$$

for each $k = 1,\cdots,\infty$.

We now have to demonstrate the existence of a sequence of unit-norm inputs $\{u_k(t)\}$ such that the corresponding outputs have norms diverging to $+\infty$. We do so by exploiting the divergence of the integral of $|h_{\alpha\beta}|$. For each $k = 1,\cdots,n$, such that:

$$u_{k,i}(t) = \begin{cases} \operatorname{sgn}\big(h_{\alpha\beta}(t_k,t)\big) \cdot \delta_{i,\beta}, & t \leq t_k \\ 0, & t > t_k \end{cases},$$

where $u_{k,i}$ denotes the i-th component of u_k, while $\delta_{i,\beta}$ is the Kronecker delta. Notice that $\|u_k(\cdot\|_\infty = 1$ for each $k \in \mathbb{N}$, and that this definition allows us to have:

$$\begin{aligned}
y_{k,\alpha}(t_k) &= \int_{-\infty}^{t_k} \sum_{i=1}^{n_i} h_{\alpha,i}(t,\tau) \cdot u_{k,i}(\tau)\,d\tau \\
&= \int_{-\infty}^{t_k} \sum_{i=1}^{n_i} \|h_{\alpha\beta}(t_k,t)\| \cdot \delta_{i,\beta}\,d\tau > k
\end{aligned}$$

By definition of $\|y_k(\cdot)\|$ as the maximum row sum of $y_k(\cdot)$:

$$\|y_k(\cdot)\| \geq |y_{k,\alpha}(t_k)| = \int_{-\infty}^{t_k} |h_{\alpha,\beta}(t_k,\tau)|d\tau > k$$

■

Corollary 4.27. *A linear time-varying system described by* (3.2) *and* (3.3)*, with bounded* $A(\cdot), B(\cdot), C(\cdot), D(\cdot)$*, is BIBO stable if and only if:*

$$\sup_{t\geq 0}\left\{\int_0^t \|C(t)\Phi(t,\tau)B(\tau)\|d\tau\right\} < \infty$$

Proof. Simply note that:

$$y(t) = \int_{-\infty}^{t} H(t,\tau)u(\tau)\,d\tau + D(t)u(t),$$

where $H(t,\tau)$ is taken here to be:

$$H(t,\tau) = \begin{cases} C(t)\Phi(t,\tau)B(\tau), & \tau \geq 0, \\ 0, & \tau < 0 \end{cases}$$

The corollary thus follows from the above theorem. ■

In particular, for linear time-invariant systems:

$$\begin{aligned} &\sup_{t\geq 0}\left\{\int_0^t \|Ce^{(t-\tau)A}B\|d\tau\right\} < \infty \\ \Leftrightarrow &\int_0^\infty \|Ce^{A\tau}B\|d\tau < \infty \end{aligned}$$

Theorem 4.28. *Consider the following transfer function:*

$$H(s) = C(sI - A)^{-1}B + D \in \mathbb{R}^{n_i \times n_o}(s)$$

Then the following is equivalent:

1. *The system is BIBO stable.*
2. $\int_0^\infty \|Ce^{tA}B\|\,dt < \infty$.
3. $\text{Poles}(H(s)) \in \mathbb{C}^o$.

Proof.

" $\Leftarrow$ " Let $G(t) \equiv Ce^{tA}B = L^{-1}\{H(s)\} - D$. Then:

$$\int_0^\infty \|Ce^{tA}B\|\, dt < \infty$$
$$\iff \int_0^\infty |g_{ij}(t)|dt < \infty,$$

Moreover, since the poles of $G(s)$ are the union of those of $g_{ij}(s)$:

$$\text{Poles}(H(s)) \subset \mathbb{C}^-$$
$$\iff \text{Poles}(G_{ij}(s)) \subset \mathbb{C}^-$$

for each $i \in \{1, \cdots, n_o\}, j \in \{1, \cdots, n_i\}$

We thus need to show that:

$$\int_0^\infty |g_{ij}(t)|dt < \infty$$
$$\implies \text{Poles}(G_{ij}(s)) \subset \mathbb{C}^-.$$

This can be shown by observing that:

$$\begin{aligned}\sup_{s\in\overline{\mathbb{C}^+}} |G_{ij}(s)| &\le \int_0^\infty \sup_{s\in\mathbb{C}^+} |g_{ij}(t)e^{-st}|\, dt \\ &\le \int_0^\infty |g_{ij}(t)|\, dt \\ &< \infty\end{aligned}$$

Thus, $|G_{ij}(s)|$ is bounded above in $\mathbb{C}^+$ by a positive constant throughout $\overline{\mathbb{C}^+}$. This demonstrates that no pole of G_{ij} lies in $\overline{\mathbb{C}^+}$; therefore, all its poles must be in $\mathbb{C}^o$.

" $\Rightarrow$ " If $\text{poles}(G(s)) \equiv \{\lambda_1, \cdots, \lambda_l\} \in \mathbb{C}^o$, then $\text{poles}(g_{ij}(t)) \in \mathbb{C}^o$. Thus, there exist polynomials $\pi_1(t), \cdots, \pi_l(t)$:

$$g_{ij}(t) = \sum_{k=1}^{l} \pi_k(t)e^{\lambda_k t},$$

For each $\epsilon > 0$, since each $\pi_k(t)$ is polynomial in t, it must grow at a slower rate than $e^{\epsilon t}$. Thus, there exists polynomials $m_1(\epsilon), \cdots, m_l(\epsilon) > 0$ such that:

$$|\pi_k(t)| \le m_k(\epsilon) \cdot e^{\epsilon t}$$

Now, define:

$$\mu \equiv \min_k\{-Re(\lambda_k)\} > 0$$
$$\epsilon \equiv \frac{1}{2}\mu$$
$$m(\epsilon) \equiv \sum_{k=1}^{l} m_k(\epsilon)$$

Thus, we have:

$$\begin{aligned}
\Rightarrow |g_{ij}(t)| &= \left| \sum_{k=1}^{l} \pi_k(t) \cdot e^{\lambda_k t} \right| \\
&\leq \sum_{k=1}^{l} |\pi_k(t)| \cdot e^{\lambda_k t} \\
&\leq \sum_{k=1}^{l} m_k(\epsilon) e^{\epsilon t} \cdot e^{-\mu t} \\
&= m(\epsilon) \cdot e^{-(\mu-\epsilon)t} \\
\Rightarrow \int_0^\infty |g_{ij}(t)|\, dt &\leq \frac{m(\epsilon)}{\mu - \epsilon} < \infty
\end{aligned}$$

Since $g(t) = Ce^{tA}B$, we have:

$$\|G(t)\| \leq \overline{m(\epsilon)} e^{-(\mu-\epsilon)t}$$

■

4.6 Lecture 14 Discussion

Example (**Discussion 9, Problem 2**). Consider the linear time-varying system:

$$\dot{x}(t) = A(t)x(t) + B(t)u(t)$$
$$y(t) = C(t)x(t) + D(t)u(t).$$

Assume that the equilibrium 0 of $\dot{x}(t) = A(t)x(t)$ is exponentially stable. Let $B(\cdot), C(\cdot), D(\cdot)$ be bounded. Show that the system is BIBO stable.

Solution:

Since $B(\cdot), C(\cdot), D(\cdot)$ are bounded, there exist $k_1, k_2, k_3 \in \mathbb{R}^+$ such that:

$$\|B(\cdot)\| \leq k_1$$
$$\|C(\cdot)\| \leq k_2$$
$$\|D(\cdot)\| \leq k_3$$

and since $\dot{x}(t) = A(t)x(t)$ is exponentially stable, there exists some $M > 0$ such that:

$$\|\Phi(t, t_0)\| \leq M \cdot e^{-\alpha(t-t_0)}$$

Given an input $u(t)$, the output $y(t)$ is:

$$\begin{aligned} y(t) &= C(t) \cdot \int_0^t \Phi(t, t_0) \cdot B(\tau)u(\tau)\, d\tau + D(t)u(t) \\ &= \cdot \int_0^t \underbrace{\left[C(t)\Phi(t, t_0)B(\tau) + D(t) \cdot \delta(t-\tau)\right]}_{\equiv H(t,\tau)} u(\tau)\, d\tau + D(t)u(t) \end{aligned}$$

$$\begin{aligned} \Rightarrow \|H(t, \tau)\| &\leq k_1 k_2 M \int_{-\infty}^t e^{-\alpha(t-t_0)}\, d\tau + k_3 \\ &= k_1 k_2 M \cdot \int_0^\infty e^{-\alpha t} dt \\ &= \frac{k_1 k_2 M}{\alpha} \end{aligned}$$

4.7 Lecture 15

Below, we describe state-space notions of stability (e.g. asymptotic stability, exponential stability), and compare these with BIBO stability. We conclude this section with Lyapunov's Lemma, which poses. This lecture draws largely from Professor Shankar Sastry's text "Nonlinear Systems: Analysis, Stability, and Control" [9]. In this lecture, the origin of each theorem, lemma, or definition originating from this text will be provided in parentheses as a convenience to the reader.

We begin our discussion with different definitions for internal stability.

Notions of Internal Stability

Below, we concern ourselves with state trajectories following the dynamics:

$$\dot{x} = f(x,t), \qquad x(t_0) = x_0 \tag{4.3}$$

Definition 4.29 (**Equilibrium Point** (Definition 5.2, pg. 184))**.** *A state $x^\star \in \Sigma$ is called an* ***equilibrium point*** *if $f(x^\star, t) = 0$ for all $t \geq 0$.*

Definition 4.30 (**Stable in the sense of Lyapunov** (Definition 5.4, pg. 185))**.**

1. *The state $x_e \equiv 0$ is called* ***stable (in the sense of Lyapunov)*** *if, for each $x_0 \in \mathbb{R}^n$ and $t_0 \in \mathbb{R}$, the mapping:*
$$x(t) = \Phi(t, t_0)x_0$$
is bounded by some positive function of t_0.

2. *The equilibrium point $x = 0$ is called a* ***stable equilibrium point*** *of the system (4.3) if, for any $t_0 \geq 0$ and $\epsilon > 0$, there exists some $\delta(t_0, \epsilon)$ such that:*
$$|x_0| < \delta(t_0, \epsilon) \quad \Rightarrow \quad |x(t)| < \epsilon, \forall t \geq t_0,$$
where $x(t)$ is the solution to (4.3), starting from $x(t_0) = x_0$.

Definition 4.31 (**Uniformly Stable** (Definition 5.5, pg. 185))**.**

1. *The state $x_e \equiv 0$ is called* ***uniformly stable*** *if, for each $x_0 \in \mathbb{R}^n$ and $t_0 \in \mathbb{R}$, the mapping:*
$$x(t) = \Phi(t, t_0)x_0$$
is bounded by some positive constant.

2. *The equilibrium point $x = 0$ is called a* ***uniformly stable equilibrium point*** *of the system if it achieves the criterion for stable equilibrium points, with some $\delta(\epsilon)$ that is independent of t_0.*

In essence, a stable (in the sense of Lyapunov) equilibrium point is uniformly stable if the associated upper bounds $\delta(t_0, \epsilon)$ for its norms never approach 0, i.e.:

$$\inf_{t_0 \in \mathbb{R}} \delta(t_0, \epsilon) > 0$$

Definition 4.32 (**Asymptotically Stable** (Definition 5.6, pg. 185-186))**.** *The state $x_e \equiv 0$ is called **asymptotically stable** if:*

1. *$x_e \equiv 0$ is a stable equilibrium point of (4.3), and*
2. *$x(t)$ converges to 0, i.e. $\lim_{t\to\infty} \Phi(t, t_0) = 0$. If this condition is met, $x = 0$ is said to be **attractive**.*

The reader may question whether it is necessary to specify the first condition "$x_e \equiv 0$ is a stable equilibrium point of (4.3)" if the second statement "$\lim_{t\to\infty} \Phi(t, t_0) = 0$" already holds true. The following example answers this question in the affirmative.

Example. Consider the dynamical system given by:

$$\begin{aligned}\dot{x}_1 &= x_1^2 - x_2^2 \\ \dot{x}_2 &= 2x_1x_2.\end{aligned}$$

The phase portrait of this system indicates that, although all trajectories following this system tends to $x = 0$ as $t \to \infty$, those particularly close to the x-axis will initially move far away from the origin before returning. In fact, one can choose a sequence of trajectories, increasingly closer to being parallel to the x-axis, such that the maximum distance (in time) between each trajectory and the origin increases as the sequence progresses. In this sense, $x = 0$ is not stable, even though it is attractive.

Definition 4.33 (**Uniformly Asymptotically Stable** (Definition 5.7, pg. 186-187))**.** *The state $x_e \equiv 0$ is called **uniformly asymptotically stable** if:*

1. *$x_e \equiv 0$ is a uniform stable equilibrium point of (4.3), and*
2. *$x(t)$ converges uniformly to 0, i.e. $\exists \delta > 0$, and $\gamma(\tau, x_0) : \mathbb{R}^+ \times \mathbb{R}^n \to \mathbb{R}^+$ such that, whenever $|x_0| < \delta$:*

$$\begin{aligned}&\|\phi(t, t_0)\| \leq \gamma(t - t_0, x_0) \\ &\lim_{\tau\to\infty} \gamma(\tau, x_0) = 0\end{aligned}$$

Let $\phi(t, x_0, t_0)$ denotes the trajectory of the system $\dot{x} = f(x, t), x(t_0) = t_0$, starting from x_0 at time t_0. Then the second condition above is equivalent to the following statement— $\exists \delta$ and some non-decreasing function $T : \mathbb{R}^+ \to \mathbb{R}^+$ such that, whenever $|x_0| < \delta$:

$$|\phi(t_1 + t, x_0, t_1)| < \epsilon$$

for each $t_1 \geq t_0$.

The definitions of asymptotic stability do not quantify the speed of convergence of trajectories to the origin, e.g. $1/t$, $1/\sqrt{t}$, etc. However, there is a particularly strong form of stability that demands an exponential rate of convergence.

Definition 4.34 (**Exponentially Stable, Rate of Convergence** (Definition 5.10, pg. 187))**.** *The state $x_e \equiv 0$ is called **exponentially stable** if $x_e \equiv 0$ is stable, and $\exists\, M, \alpha > 0$ such that:*

$$\|x(t)\| \leq Me^{-\alpha(t-t_0)} \cdot |x_0|$$

We will later show that, for linear systems (whether time-invariant or time-varying), uniform asymptotic stability implies exponential stability. However, we first examine a few basic results regarding asymptotic stability.

Theorem 4.35 (**Asymptotic Stability** (Theorem 5.33, pg. 209))**.** *For a time-invariant system, $x = 0$ is asymptoptically stable if and only if, for each $t_0 \in \mathbb{R}$:*

$$\lim_{t\to\infty} \Phi(t, 0) = O$$

Remark. Since $t \to \Phi(t, 0)$ is continuous, that $\lim_{t\to\infty} \phi(t, 0) \to 0$ implies the boundedness of $\Phi(t, t_0)$ in t, for each $t_0 \in \mathbb{R}$.

Proof.

" $\Rightarrow$ " The proof is most readily established by contradiction. Suppose that $\Phi(t, 0)$ does not approach 0 as $t \to \infty$. Then there must exist some $i, j \in \{1, \cdots, n\}$ such that:

$$\phi_{i,j}(t, 0) \not\to O$$

as $t \to \infty$. Now, choose x_0 to be the j-th standard vector in $\mathbb{R}^n$, i.e. with one for the j-th element and zero for all other elements. Then the i-th component of $x(t) = \phi(t, 0)x_0$ will not approach 0 as $t \to \infty$, contradicting the fact that $x = 0$ is asymptotically stable.

" $\Leftarrow$ " The proof can be completed by establishing bounds on $\phi(t, 0)$ and $x(t)$. Fix some initial state x_0 and some initial time t_0, and let $\epsilon > 0$. Since $\phi(t, 0) \to O$ as $t \to \infty$, there exists some $t_M \in \mathbb{R}^+$ such that:

$$\|\Phi(t, 0)\| < \frac{\epsilon}{|x_0| \cdot \|\Phi(0, t_0)}$$

whenever $t > t_M$. Then, when $t > t_M$:

$$|x(t)| = |\phi(t, 0)x_0| \leq \|\Phi(t, 0)\| \cdot \Phi(0, t_0) \cdot |x_0| < \epsilon$$

This establishes the asymptotic stability of $x = 0$.

■

Lemma 4.36 (**Exponential Stability of $\dot{x} = Ax$**)**.** *The system $\dot{x} = Ax$ is exponentially stable if and only if:*

$$\sigma(A) \subset \mathbb{C}^-$$

Proof. Applying the Jordan form of A and the Cayley-Hamilton Theorem, there exist an invertible matrix P, and polynomials $\pi_1(t), \cdots, \pi_n(t)$, of order less than n (constructed from the elements of P) such that:

$$e^{At} = Pe^{tJ}P^{-1} = \sum_{k=1}^{n} \pi_k(t) \cdot e^{\lambda_k t},$$

where J is the Jordan form of A.

" $\Rightarrow$ " Again, the proof is most readily established by contradiction. If at least one of the eigenvalues of A is not in $\mathbb{C}^-$, then $e^{At} = \sum_{k=1}^{n} \pi_k(t) \cdot e^{\lambda_k t}$ does not tend to 0 as $t \to \infty$. Thus, $\dot{x} = Ax$ is *not* exponentially stable.

" $\Leftarrow$ " Since $\sigma(A) = \{\lambda_k | k = 1, \cdots, n\} \in \mathbb{C}^-$ by hypothesis, and the behavior of exponential functions dominate that of polynomial functions over time, it follows that $\dot{x} = Ax$ is exponentially stable.

■

The following theorem summarizes the relationship between the internal stability of the system $\dot{x} = Ax$ and the Jordan decomposition of A.

Theorem 4.37. *Let $A \in \mathbb{R}^n$ be given, and let $\sigma(A)$ denote the set of all eigenvalues of A, i.e. the spectrum of A. Then the system $\dot{x} = Ax$ is internally stable if and only if both of the following conditions hold:*

1. $\sigma(A) \subset \overline{\mathbb{C}^-}$
2. *The Jordan blocks for each $\lambda \in \sigma(A) \cap \mathbb{C}^o$ are all of size 1.*

Proof. We provide a succinct, though not completely mathematically rigorous, explanation.

Let $\sigma(A) = \{\lambda_1, \cdots, \lambda_k\}$, where $k \leq n$. Let a_i, g_i and m_i denote the algebraic multiplicity, geometric multiplicity, and index for each $i = 1, \cdots, k$. Let:

$$A = PJP^{-1}$$

be the Jordan decomposition of A. Recall that the stability of the solution, $x(t) = e^{tA}x(0)$, depends on the time dependence of the matrix exponential e^{tA}.

First, since the Jordan decomposition of A is block-diagnol, $\mathbb{R}^n$ can be written as the direct sum of a collection of subspaces, each of which is spanned by a particular Jordan chain consisting of an eigenvector of A and generalized eigenvectors derived from that eigenvector. More precisely, by recalling that the geometric multiplicity of each eigenvalue gives the number of Jordan blocks associated with that eigenvalue:

$$\begin{aligned} \mathbb{R}^n &= (V_{\lambda_1,1} \oplus \cdots \oplus V_{\lambda_1,g_1}) \oplus \cdots \oplus (V_{\lambda_k,1} \oplus \cdots \oplus V_{\lambda_k,g_k}) \\ \Rightarrow\ J &= \text{diag}\{J_{\lambda_1,1}, \cdots, J_{\lambda_1,g_1}, \cdots, J_{\lambda_k,1}, \cdots, J_{\lambda_k,g_\sigma}\} \end{aligned}$$

Clearly, $x(t) = e^{tA}x(0)$ is stable if and only if e^{tA} contains only constant or exponentially decaying terms; this, in turn, is true for e^{tA} if and only if it is true for each $e^{tJ_{\lambda_i,j}}$, where $i = 1, \cdots, \sigma, j = 1, \cdots, g_i$. If $\sigma(A) \subset \mathbb{C}^-$, this is always true; if there exists some eigenvalue on the imaginary axis, i.e. $\exists \lambda \in \sigma(A) \cap \mathbb{C}^o$, then we require that eigenvalue to have Jordan blocks of size 1. This is because Jordan blocks of sizes larger than 1 would involve polynomial terms in t, which, for eigenvalues of real part 0, would result in divergence in off-diagonal terms of the Jordan block as $t \to \infty$. ■

Theorem 4.38 (**Exponential and Uniform Asymptotic Stability** (Theorem 5.33, pg. 209))**.** *Consider a linear, time-varying system of the form:*

$$\dot{x} = A(t)x, \qquad , x(t_0) = x_0,$$

where $A(t)$ is piecewise continuous and bounded. Then the following statements are equivalent:

1. *$x = 0$ is a uniform asymptotic stable equilibrium point of this system.*
2. *$x = 0$ is an exponentially stable equilibrium point of this system.*

Remark. The "if" direction is trivial. Our strategy for the "only if" direction will be to show that, for any fixed t_0, as t increase in time, the norm of the state transition matrix $\Phi(t, t_0)$ must decay in a geometric series, whenever t increases by some T.

Proof. By definition, exponential stability implies uniform asymptotic stability. For the converse, suppose the system is uniformly asymptotically stable. Fix some $t_0 \geq 0$; then, for each $t_1 \geq 0$, there exists some $m_0 > 0$ such that:

$$\|\Phi(t, t_1)\| \leq m_0, \qquad t \geq t_1$$

Next, the uniform convergence of $\Phi(t, t_0)$, (which follows from the uniform asymptotic stability of $x = 0$) implies that there exists some $T > 0$ for which:

$$\|\Phi(t, t_1)\| < \frac{1}{2},$$

whenever $t - t_1 > T$. Now, fix $t > t_0$, and let $k \in \{0, 1, 2, \cdots, \}$ be given such that:

$$t_0 + (k-1)T \leq t \leq t_0 + kT$$

Then, we have:

$$\begin{aligned}
\|\Phi(t, t_0)\| &= \left| \Phi(t, t_0 + kT) \cdot \prod_{j=1}^{k} \Phi\big(t_0 + jT, t_0 + (j-1)T\big) \right| \\
&\leq \|\Phi(t, t_0 + kT)\| \cdot \prod_{j=1}^{k} \|\Phi\big(t_0 + jT, t_0 + (j-1)T\big)\| \\
&\leq m_0 \cdot 2^{-k} \\
&\leq m_0 \cdot 2^{-\frac{t-t_0}{T}},
\end{aligned}$$

since $k \geq \frac{t-t_0}{T}$. This establishes the exponential decay of $\|\Phi(t, t_0)\|$ towards 0 as $t \to \infty$. ■

The following theorem provides a method of determining whether the eigenvalues of A all lie in the left half plane that is computationally more efficient than explicitly solving for the n roots of the n-degree polynomial $\chi_A(s)$.

Theorem 4.39 (**Routh-Hurwitz Criterion for Matrix Stability**). *Let $A \in \mathbb{R}^{n\times n}$, and suppose the characteristic polynomial of A is:*

$$\chi_A(s) = s^n + a_{n-1}s^{n-1} + \cdots + a_1 s + a_0,$$

where $a_0, \cdots, a_{n-1} \in \mathbb{R}$. Consider the matrices:

$$\begin{aligned}
D_1 &= a_{n-1}, \\
D_2 &= \begin{bmatrix} a_{n-1} & a_{n-3} \\ a_n & a_{n-2} \end{bmatrix}, \\
D_3 &= \begin{bmatrix} a_{n-1} & a_{n-3} & a_{n-5} \\ a_n & a_{n-2} & a_{n-4} \\ 0 & a_{n-1} & a_{n-3} \end{bmatrix} \\
&\vdots \\
D_n &= \begin{bmatrix} a_{n-1} & a_{n-3} & a_{n-5} & \cdots & \cdots & 0 \\ a_n & a_{n-2} & a_{n-4} & \cdots & \cdots & 0 \\ 0 & a_{n-1} & a_{n-3} & \cdots & \cdots & 0 \\ 0 & a_n & a_{n-2} & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \cdots & a_1 & 0 \\ 0 & 0 & \cdots & \cdots & a_2 & a_0 \end{bmatrix}
\end{aligned}$$

Then $\sigma(A) \subset \mathbb{C}^-$ if and only if $\det(D_i) > 0$ for each $i = 1, \cdots, n$. This is known as the Hurwitz criterion.

Example. Using the Hurwitz condition, find the conditions for α under which the system governed by:

$$\ddot{y} + 2\alpha\dot{y} + y = 0$$

is stable.

We conclude this section by noting that, although exponential stability in linear systems implies BIBO stability, the converse is, in general, not true.

Theorem 4.40. *Consider the linear system:*

$$\begin{aligned}
\dot{x}(t) &= A(t)x + B(t)u \\
y(t) &= C(t)x + D(t)u
\end{aligned}$$

If $x = 0$ be an exponentially stable point of this system, and $B(\cdot), C(\cdot), D(\cdot)$ are bounded, the system is BIBO stable.

Proof. Since $B(\cdot), C(\cdot), D(\cdot)$ are bounded, there exist $k_1, k_2, k_3 > 0$ such that:

$$\begin{aligned}\|C(\cdot)\| &\leq k_1,\\ \|B(\cdot)\| &\leq k_2,\\ \|D(\cdot)\| &\leq k_3\end{aligned}$$

The zero-state response of the system is:

$$y(t) = \int_0^t C(t)\Phi(t,\tau)B(\tau)u(\tau)\,d\tau + D(t)u(t)$$

Thus, we need to show that:

$$\begin{aligned}&\sup_{t\in\mathbb{R}} \int_{-\infty}^t \|C(t)\Phi(t,\tau)B(\tau) + D(t)\delta(t-\tau)\|\,d\tau\\ \leq &\sup_{t\in\mathbb{R}} \int_{-\infty}^t \|C(t)\| \cdot \|\Phi(t,\tau)\| \cdot \|B(\tau)\|\,d\tau + \|D(t)\|\\ \leq &\sup_{t\in\mathbb{R}} \int_{-\infty}^t k_1 \cdot me^{-\mu(t-\tau)} \cdot k_2\,d\tau\\ = &\frac{mk_1k_2}{\mu} < \infty\end{aligned}$$

This establishes the BIBO stability of the given system. ■

BIBO vs. Internal Stability, Stability in Time-Variant Systems

The converse to the above theorem is not true, i.e. BIBO-stable linear systems are not necessarily exponentially stable, or even internally stable. It is not true even for time-invariant systems. This is because the "output matrix" C and "control matrix" C may be chosen such that the internal instability of the system, characterized by eigenvalues of A in $\mathbb{C}^+$ (or possibly, on $\mathbb{C}^o$), does not appear in the transfer function:

$$H(s) = C(sI - A)^{-1}B + D$$

If this occurs for some choice of C, the unstable modes of the system are said to be *unobservable*; if this occurs for some choice of B, the unstable modes of the system are said to be *uncontrollable*. This is illustrated by the inverted pendulum example following Lecture 10, and in the following examples.

This concept can be explained in a slightly different manner. In bygone eras of control theory, engineers may be given a transfer function $H(s)$ that models some system, and asked to design a linear system to implement $H(s)$. Mathematically speaking, the task at hand would be to find suitable state and input vectors, $x(t)$ and $u(t)$, respectively, and matrices A, B, C, D such that the system:

$$\begin{aligned}\dot{x}(t) &= A(t)x(t) + B(t)u(t)\\ y(t) &= C(t)x(t) + D(t)u(t)\end{aligned}$$

has the transfer function $H(s)$. If the system can be made time-invariant, the latter task boils down to finding A, B, C, D such that:

$$H(s) = C(sI - A)^{-1}B + D$$

Choices of A, B, C, D that implement the transfer function $H(s)$ are called *realizations* of $H(s)$.

It is easy to see that realizations for a given transfer function are not unique. Given a realization, one could, for example, multiply each entry of C by 2 and divide each entry of D by 2 without changing $H(s)$. One could also choose B, C such that certain poles of A do not appear in the transfer function $H(s) = C(sI - A)^{-1}B + D$. Of particular interest are choices of $x(t)$, $u(t)$, and A, B, C, D such that all the modes of A are controllable and observable; such a realization is said to be *minimal*. Minimal realizations force all the eigenvalues of A to appear as modes in the transfer function $H(s)$, and thus eliminates the asymmetry between BIBO stability and exponential stability. In other words, for a minimally realized linear time-invariant system S, the following are equivalent:

- $\text{Poles}(H(s)) = \sigma(A) \subset \mathbb{C}^-$
- S is exponentially stable.
- S is BIBO stable.

As the following examples illustrate that this equivalence does not hold for realizations in general.

The figure on the next page summarizes the relationship between BIBO and internal stability.

Example (**Lecture 15, pg. 8, Example 1**). Consider the linear dynamical system:

$$\begin{cases} \dot{x} &= Ax + Bu \\ y &= Cx, \end{cases}$$

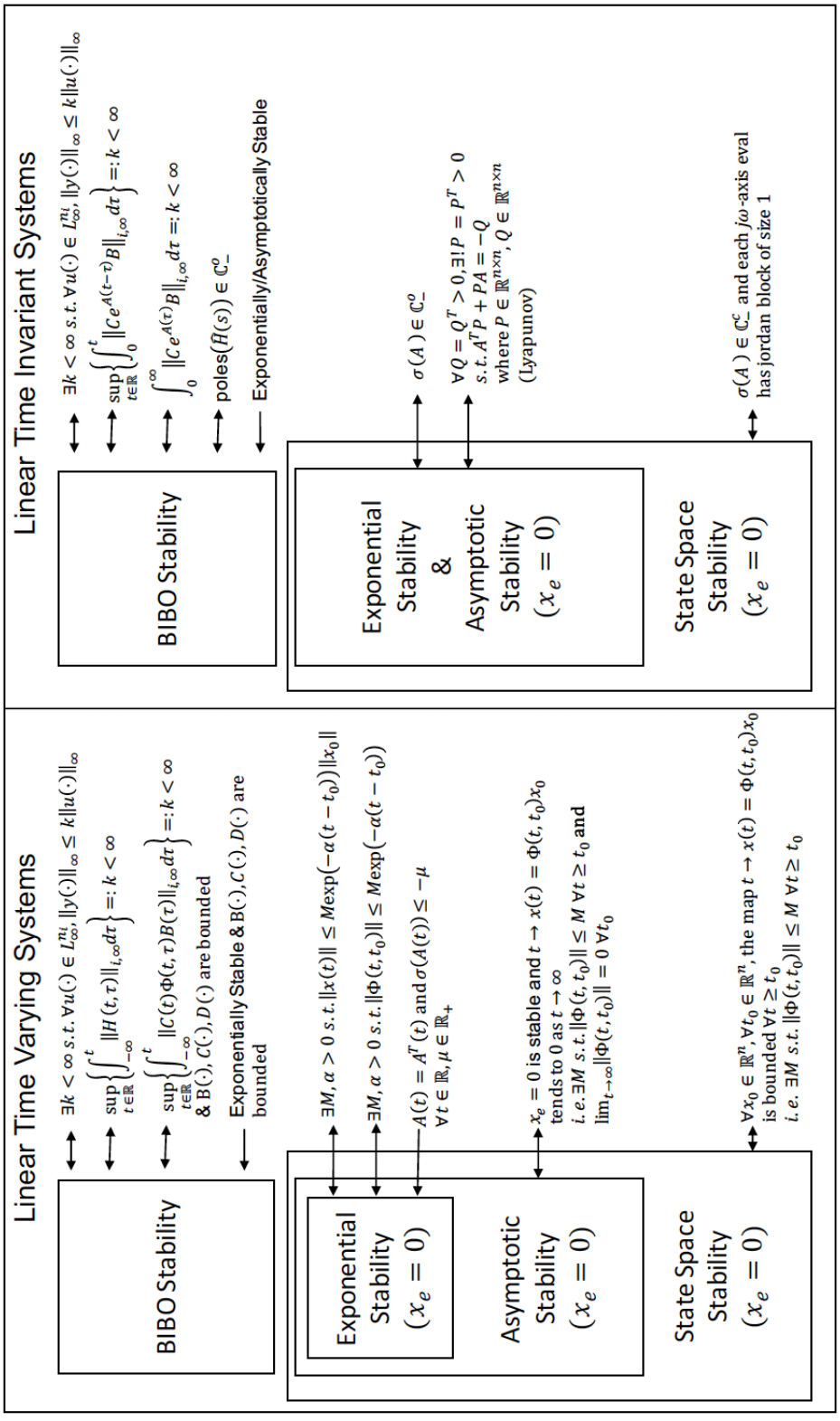
where:

$$A = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -3 & 0 \\ 0 & 0 & -6 \end{bmatrix}, \qquad B = \begin{bmatrix} 1/10 \\ -1/6 \\ 1/15 \end{bmatrix}, \qquad C = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}$$

Discuss the internal and BIBO stability of this system.

Solution:

Since $\sigma(A) = \{-1, -3, -6\} \subset \mathbb{C}^-$, the given system is internally stable. Thus, the transfer function $G(s) = C(sI - A)^{-1}B$ can only have roots at $s = -1, -3, -6$. As such, the given system is also BIBO stable.


Linear Time Varying Systems
BIBO Stability
$\exists k < \infty \; s.t. \; \forall u(\cdot) \in L^{n_i}_\infty, \|y(\cdot)\|_\infty \le k\|u(\cdot)\|_\infty$
$\sup_{t\in\mathbb{R}} \left\{ \int_{-\infty}^{t} \|H(t,\tau)\|_{i,\infty} d\tau \right\} =: k < \infty$
$\sup_{t\in\mathbb{R}} \left\{ \int_{-\infty}^{t} \|C(t)\Phi(t,\tau)B(\tau)\|_{i,\infty} d\tau \right\} =: k < \infty$
& B(·), C(·), D(·) are bounded
Exponentially Stable & B(·), C(·), D(·) are bounded
Exponential Stability ($x_e = 0$)
$\exists M, \alpha > 0 \; s.t. \|x(t)\| \le M\exp(-\alpha(t-t_0))\|x_0\|$
$\exists M, \alpha > 0 \; s.t. \|\Phi(t,t_0)\| \le M\exp(-\alpha(t-t_0))$
$A(t) = A^T(t)$ and $\sigma(A(t)) \le -\mu$
$\forall t \in \mathbb{R}, \mu \in \mathbb{R}_+$
Asymptotic Stability ($x_e = 0$)
$x_e = 0$ is stable and $t \to x(t) = \Phi(t,t_0)x_0$ tends to 0 as $t \to \infty$
i.e. $\exists M \; s.t. \|\Phi(t,t_0)\| \le M \; \forall t \ge t_0$ and $\lim_{t\to\infty}\|\Phi(t,t_0)\| = 0 \; \forall t_0$
State Space Stability ($x_e = 0$)
$\forall x_0 \in \mathbb{R}^n, \forall t_0 \in \mathbb{R}^n$, the map $t \to x(t) = \Phi(t,t_0)x_0$ is bounded $\forall t \ge t_0$
i.e. $\exists M \; s.t. \|\Phi(t,t_0)\| \le M \; \forall t \ge t_0$
Linear Time Invariant Systems
BIBO Stability
$\exists k < \infty \; s.t. \; \forall u(\cdot) \in L^{n_i}_\infty, \|y(\cdot)\|_\infty \le k\|u(\cdot)\|_\infty$
$\sup_{t\in\mathbb{R}} \left\{ \int_{0}^{t} \|Ce^{A(t-\tau)}B\|_{i,\infty} d\tau \right\} =: k < \infty$
$\int_0^\infty \|Ce^{A(\tau)}B\|_{i,\infty} d\tau =: k < \infty$
poles$(H(s)) \in \mathbb{C}^o_-$
Exponentially/Asymptotically Stable
Exponential Stability & Asymptotic Stability ($x_e = 0$)
$\sigma(A) \in \mathbb{C}^o_-$
$\forall Q = Q^T > 0, \exists! P = P^T > 0$ $s.t. A^TP + PA = -Q$ where $P \in \mathbb{R}^{n\times n}$, $Q \in \mathbb{R}^{n\times n}$ (Lyapunov)
State Space Stability ($x_e = 0$)
$\sigma(A) \in \mathbb{C}^c_-$ and each $j\omega$-axis eval has jordan block of size 1

More explicitly, we have for the transfer function:

$$\begin{aligned} G(s) &= C(sI - A)^{-1}B \\ &= \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{s+1} & 0 & 0 \\ 0 & \frac{1}{s+3} & 0 \\ 0 & 0 & \frac{1}{s+6} \end{bmatrix} \begin{bmatrix} 1/10 \\ 0 \\ 1/15 \end{bmatrix} \\ &= \frac{1}{10} \cdot \frac{1}{s+1} - \frac{1}{6} \cdot \frac{1}{s+3} + \frac{1}{15} \cdot \frac{1}{s+6} \\ &= \frac{1}{(s+1)(s+3)(s+6)} \end{aligned}$$

Thus, the given system is BIBO stable.

Example (**Lecture 15, pg. 8, Example 2**). Consider the linear dynamical system:

$$\begin{cases} \dot{x} &= Ax + Bu \\ y &= Cx, \end{cases}$$

where:

$$A = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & -6 \end{bmatrix}, \qquad B = \begin{bmatrix} 1/10 \\ 0 \\ 1/15 \end{bmatrix}, \qquad C = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}$$

Discuss the internal and BIBO stability of this system.

Solution:

Since $\sigma(A) = \{-1, 3, -6\} \not\subset \overline{\mathbb{C}^-}$, the given system is not internally stable.

To consider the BIBO stability of the system, consider the transfer function:

$$\begin{aligned} G(s) &= C(sI - A)^{-1}B \\ &= \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{s+1} & 0 & 0 \\ 0 & \frac{1}{s-3} & 0 \\ 0 & 0 & \frac{1}{s+6} \end{bmatrix} \begin{bmatrix} 1/10 \\ 0 \\ 1/15 \end{bmatrix} \\ &= \frac{1}{10} \cdot \frac{1}{s+1} - \frac{1}{6} \cdot \frac{1}{s+3} + \frac{1}{15} \cdot \frac{1}{s+6} \\ &= \frac{1}{6} \cdot \frac{s+4}{(s+1)(s+6)} \end{aligned}$$

Thus, the given system is BIBO stable, despite not being *internally* stable. This is because, the mode $\lambda = 3$ is uncontrollable from u, i.e. the corresponding pole disappears during the derivation of the transfer function when $(sI - A)^{-1}$ is multiplied to the right by B.

Example (**Lecture 15, pg. 11, Example 3**). Consider the linear dynamical system:

$$\begin{cases} \dot{x} &= Ax + Bu \\ y &= Cx, \end{cases}$$

where:

$$A = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & -6 \end{bmatrix}, \qquad B = \begin{bmatrix} 1/10 \\ -1/6 \\ 1/15 \end{bmatrix}, \qquad C = \begin{bmatrix} 1 & 0 & 1 \end{bmatrix}$$

Discuss the internal and BIBO stability of this system.

Solution:

Since $\sigma(A) = \{-1, 3, -6\} \not\subset \overline{\mathbb{C}^-}$, the given system is not internally stable.

To consider the BIBO stability of the system, consider the transfer function:

$$\begin{aligned} G(s) &= C(sI - A)^{-1}B \\ &= \begin{bmatrix} 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{s+1} & 0 & 0 \\ 0 & \frac{1}{s-3} & 0 \\ 0 & 0 & \frac{1}{s+6} \end{bmatrix} \begin{bmatrix} 1/10 \\ -1/6 \\ 1/15 \end{bmatrix} \\ &= \frac{1}{10} \cdot \frac{1}{s+1} - \frac{1}{6} \cdot \frac{1}{s+3} + \frac{1}{15} \cdot \frac{1}{s+6} \\ &= \frac{1}{6} \cdot \frac{s+4}{(s+1)(s+6)} \end{aligned}$$

Thus, the given system is BIBO stable, despite not being *internally* stable. This is because, the mode $\lambda = 3$ is unobservable from y, i.e. the corresponding pole disappears during the derivation of the transfer function when $(sI - A)^{-1}$ is multiplied to the left by C.

For time-varying system, there exists no connection between the eigenvalues of $A(t)$ and stability. Even if $\sigma(A(t)) = \{-1\} \in \mathbb{C}^-$, the matrix $A(t)$ may contain time-varying terms that cause the state $x(t)$ to become unbounded as $t \to \infty$. As an example, consider the system shown below.

Example. Consider the system characterized by:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} -1 & e^{2t} \\ 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \qquad x_0 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Solving the above differential equations, we have:

$$\begin{aligned} x_2 &= e^{-t} \\ \Rightarrow \dot{x}_1 &= -x_1 + e^{2t} \cdot x_2 \\ &= -x_1 + e^t \\ \Rightarrow x_1(t) &= \frac{1}{2}(e^t - e^{-t}). \end{aligned}$$

In short, due to the off-diagonal term e^{2t} in $A(t)$, we have $x_1(t) \to \infty$ as $t \to \infty$, despite the fact that $\sigma(A(t)) = \{-1\} \in \mathbb{C}^-$.

However, there are in fact two notable cases when the eigenvalues of $A(t)$ reveal a significant amount of information regarding the stability of a system.

The first case arises when $A(t) = A^T(t)$ (i.e. when A is symmetric).

Theorem 4.41. *If $A(t)$ is symmetric, and there exists some $\mu > 0$ such that each eigenvalue of $A(t)$ is no greater than $-\mu$, then 0 is exponentially stable.*

Proof. Given the system $\dot{x} = Ax$, and any initial point $x(0)$, consider a simple energy function on the corresponding trajectory:

$$x^T x = |x|^2$$

(Energy functions of more complex forms will be given in subsequent subsections of this lecture). Then:

$$\begin{aligned} \frac{d}{dt}(|x|^2) &= \dot{x}^T x + x^T \dot{x} \\ &= x^T A^T(t) x + x^T A(t) x \\ &= 2x^T A(t) x \\ &\leq -2\mu x^T x \\ \Rightarrow |x(t)|^2 &\leq |x(0)|^2 \cdot e^{-2\mu t} \end{aligned}$$

This shows that $|x(t)| \leq \|x(0)\| \cdot e^{-\mu t}$, which establishes the exponential stability of the system. ■

Theorem 4.42. *If there exist some $\lambda > 0$ and some sufficiently small $\epsilon > 0$ such that $A(t)$ satisfies:*

$$Re(A(t)) \leq -\lambda < 0$$

for each $t \in \mathbb{R}$, and $\|A(t)\| \leq \epsilon$, then the system is stable.

Proof. (Beyond the scope of this course). ■

Stability and Energy Functions:

The Basic Stability Theorem of Lyapunov, presented below, illustrates that the different definitions of stability mentioned above can be directly characterized by an energy function $V(x,t)$ that describes the system. This energy function is often upper and/or lower bounded by a set of continuous functions with particular properties. We first present definitions of broad classes of functions that satisfy these properties.

Definition 4.43 (**Classes of Functions, Part 1** (Definition 5.12, pg. 188))**.**

1. *A function $\alpha(\cdot) : \overline{\mathbb{R}^+} \to \overline{\mathbb{R}^+}$ belongs to **class K**, denoted by $\alpha(\cdot) \in K$, if it is continuous, strictly increasing, and $\alpha(0) = 0$.*

2. *A function $\alpha(\cdot) : \overline{\mathbb{R}^+} \to \overline{\mathbb{R}^+}$ belongs to **class KR**, denoted by $\alpha(\cdot) \in K$, if $\alpha \in K$ and $\alpha(p) \to \infty$ as $p \to \infty$.*

Below, we characterize functions that behave locally and globally "like an energy function," in the sense that they increase in the direction away from the origin (which can, in the context of these definitions, be intuitively thought of as an attractive equilibrium point).

Definition 4.44 (**Classes of Functions, Part 2** (Definitions 5.12, 5.13, 5.14, pg. 188))**.**

1. *A function* $v(x,t) : \mathbb{R}^n \times \overline{\mathbb{R}^+} \to \overline{\mathbb{R}^+}$ *is called* ***locally positive definite (l.p.d.)*** *if it is continuous, and there exists some* $h > 0$ *and some function* $\alpha(\cdot) \in K$ *such that:*

$$\begin{aligned} &v(0,t) = 0, \\ &v(x,t) \geq \alpha(|x|), \qquad \forall\, x \in B_h, \qquad t \geq 0 \end{aligned}$$

2. *A function* $v(x,t) : \mathbb{R}^n \times \overline{\mathbb{R}^+} \to \overline{\mathbb{R}^+}$ *is called* ***(globally) positive definite (p.d.)*** *if it is continuous, and there exists some function* $\alpha(\cdot) \in KR$ *such that:*

$$\begin{aligned} &v(0,t) = 0, \\ &v(x,t) \geq \alpha(|x|), \qquad \forall\, x \in \mathbb{R}^n, \qquad t \geq 0 \end{aligned}$$

3. *A function* $v(x,t) \to \mathbb{R}^n \times \overline{\mathbb{R}^+} \to \overline{\mathbb{R}^+}$ *is called* ***decrescent*** *if it is continuous, and there exists some function* $\beta(\cdot) \in K$ *such that:*

$$v(x,t) \leq \beta(|x|), \qquad \forall\, x \in B_h, \qquad t \geq 0$$

Remark. If $v(x,t)$ does not explicitly depend on the time t, it must be decrescent. This is because it is either bounded above by a function of class K, or unbounded above, in which case it is bounded by itself. In addition, if $v(x,t)$ is decrescent, then $v(0,t) \leq \beta(0) = 0$. (The equality follows from $\beta(\cdot) \in K$).

Examples are given below for each of the above types of functions.

Example (**Examples of l.p.d., p.d., and decrescent functions** (Example 5.15, pgs. 188-189)). Here are some examples of energy-like functions and their membership in the various classes introduced above. It is an interesting exercise to check the appropriate functions of class K and KR that can to be used to verify these properties.

For the examples below, P is positive definite, whereas Q is not. No other information is assumed about P or Q.

Table 4.1: Classification of Value Functions

	$v(x,t)$	l.p.d.f.	p.d.f.	Decrescent
(1)	$\lvert x^2 \rvert$	Yes	Yes	Yes
(2)	$x^T P x$	Yes	Yes	Yes
(3)	$(t+1)\lvert x \rvert^2$	Yes	Yes	No
(4)	$e^{-t}\lvert x \rvert^2$	No	No	Yes
(5)	$\sin^2(\lvert x \rvert^2)$	Yes	No	Yes
(6)	$e^t x^T Q x$	No	No	No

The theorem below illustrates how imposing an increasingly strict set of conditions on the value function $v(x,t)$ and its time derivative $\dot{v}(x,t)$, defined *along the trajectory of the system's state*, allows us to make increasingly stronger claims regarding the stability of the system. In particular, we define $\dot{v}(x,t)$ as:

$$\begin{aligned}\frac{dv}{dt}(x,t)\bigg|_{\substack{\dot{x}=f(x,t)\\ x(t_0)=x_0}} &= \frac{\partial v}{\partial t}(x,t) + \frac{\partial v}{\partial x}(x,t)\frac{dx}{dt} \\ &= \frac{\partial v}{\partial t}(x,t) + \frac{\partial v}{\partial x}(x,t)f(x,t)\end{aligned}$$

This is called the *Lie derivative* of $v(x,t)$ *along* $f(x,t)$.

Theorem 4.45 (**Basic Lyapunov Theorems** (Theorem 5.16, pg. 189))**.** *Sets of conditions on $v(x,t)$ and $\dot{v}(x,t)$ are associated with notions of internal stability as given in the following table. Without loss of generality, we have placed the equilibrium point at the origin.*

Table 4.1

Table 4.2: Basic Lyapunov Theorems

	Conditions on $v(x,t)$	Conditions on $-\dot{v}(x,t)$	Conclusions
1	l.p.d.f.	≥ 0 locally	stable
2	l.p.d.f., decrescent	≥ 0 locally	unif. stable
3	l.p.d.f., decrescent	l.p.d.f.	unif. asymp. stable
4	p.d.f., decrescent	p.d.f.	globally unif. asymp. stable

Proof. (see Appendix) ■

Next, we wish to examine the stability of exponential functions.

Theorem 4.46 (**Exponential Stability Theorem** (Theorem 5.17, pg. 195))**.** *Suppose $f(x,t) : \mathbb{R}^n \times \mathbb{R}^+ \to \mathbb{R}^n$ has continuous first partial derivatives in x, and is piecewise continuous in t. Then the following two statements are equivalent:*

1. *$x = 0$ is a locally exponentially stable equilibrium point of $\dot{x} = f(x,t)$:, i.e. there exists some $h, m, \alpha > 0$ such that for each $x \in B_h$:*

$$|\Phi(t,t_0)| \leq me^{-\alpha(t-t_0)}$$

2. *There exists a function $v(x,t)$ and some $h, \alpha_1, \alpha_2, \alpha_3, \alpha_4 > 0$ such that:*

$$\begin{aligned}\alpha_1|x|^2 \leq v(x,t) &\leq \alpha_2|x|^2 \\ \frac{dv}{dt}(x,t)\bigg|_{\substack{\dot{x}=f(x,t)\\ x(t_0)=x_0}} &\leq -\alpha_3|x|^2 \\ \left|\frac{\partial v}{\partial x}(x,t)\right| &\leq \alpha_4|x|\end{aligned}$$

Proof.
"(1) $\Rightarrow$ (2)" : (see Appendix).

"(2) $\Rightarrow$ (1)" : This direction is straightforward. Note that:

$$\begin{aligned}
&\dot{v}(x,t) \leq -\frac{\alpha_3}{\alpha_2} v(x,t), \\
&\Rightarrow v(x(t),t) \leq v(x(t_0),t_0) \cdot e^{-\frac{\alpha_3}{\alpha_2}(t-t_0)} \\
&\Rightarrow \alpha_1 |x(t)|^2 \leq v(x(t),t) \leq v(x(t_0),t_0) \cdot e^{-\frac{\alpha_3}{\alpha_2}(t-t_0)} \leq \alpha_2 |x(t_0)|^2 e^{-\frac{\alpha_3}{\alpha_2}(t-t_0)} \\
&\Rightarrow |x(t) \leq \sqrt{\frac{\alpha_2}{\alpha_1}} |x(t_0)| \cdot e^{-\frac{\alpha_3}{\alpha_2}(t-t_0)}
\end{aligned}$$

■

Lyapunov Equations:

Below, we motivate the definition of Lyapunov equations. Consider the time-varying system:

$$\dot{x} = A(t)x$$

and suppose this system is associated with some notion of stability that is characterized by a potential function:

$$V(x,t) \equiv x^\star P(t) x$$

Intuitively, if the system is stable, then the potential decreases in time, i.e.:

$$\begin{aligned}
0 \geq \frac{dV}{dt} &= \dot{x}^\star P(t)x + x^\star P(t)\dot{x} + x^\star \dot{P}(t)x \\
&= (x^\star A(t)^\star) P(t)x + x^\star P(t)(A(t)x) + x^\star \dot{P}(t)x \\
&= x^\star (\underbrace{A(t)^\star P(t) + P(t)A^\star(t) + \dot{P}(t)}_{\equiv -Q(t)})x
\end{aligned}$$

In other words, we want:

$$A(t)^\star P(t) + P(t)A^\star(t) + \dot{P}(t) = -Q(t)$$

with $Q(t)$ positive definite.

The above concepts leads to Lyapunov's Lemma, presented below. However, we first present the following lemma to illustrate the connection between the Lyapunov equation and exponential stability. In a way, this is a more generalized version of Lyapunov's Lemma.

Lemma 4.47 (**Time-Varying Lyapunov Lemma** (Claim 5.38, Theorem 5.40, pgs. 212-213))**.** *Given a system $\dot{x} = A(t)x$ with state transition matrix $\Phi(t,t_0)$, the following statements are equivalent:*

1. *The system $\dot{x} = A(t)x$ is exponentially stable.*
2. *There exist constants $\alpha, \beta, \mu > 0$, and some positive definite $Q(t)$ such that the following two conditions hold:*
   - *$Q(t) \geq \mu I$, and:*
   - *The Lypaunov equation $\dot{P} + A^\star P + PA = -Q$ has a solution $P(t)$:*
   $$P(t) = \int_t^\infty \Phi^\star(\tau, t)\, Q(\tau)\, \Phi(\tau, t)\, d\tau$$
   satisfying $\alpha I \leq P(t) \leq \beta I$.

Proof. For both sides of the proof, we must establish that the function:

$$P(t) = \int_t^\infty \Phi^\star(\tau, t)\, Q(\tau)\, \Phi(\tau, t)\, d\tau$$

solves the Lyapunov equation $\dot{P} + A^\star P + PA = -Q$. To begin with, it is uncertain whether or not $P(t)$, as given by the above improper integral, is even well-defined; that would depend on properties of $\Phi(\tau, t)$, to be verified in the "(1) $\Rightarrow$ (2)" portion of the proof. If, however, $P(t)$ is indeed well defined, we can directly differentiate $P(t)$ via differentiation under the integral sign:

$$\begin{aligned}
&\frac{d}{dt}\left(\int_t^\infty \Phi^\star(\tau, t)\, Q(\tau)\, Q(\tau, t)\, d\tau\right) \\
&= -Q(t) + \int_t^\infty \left[A^\star(t)\Phi^\star(\tau, t)\, Q(\tau)\, Q(\tau, t) + \Phi^\star(\tau, t)\, Q(\tau)\, Q(\tau, t)A(t)\right] d\tau \\
&= -Q(t) + A^\star(t)P(t) + P(t)A(t)
\end{aligned}$$

Thus, *if $P(t)$ is well-defined*, it satisfies the Lyapunov equation $\dot{P} + A^\star P + PA = -Q$.

"(2) $\Rightarrow$ (1)" : By the Basic Lyapunov Theorems, it suffices to show the following:

- $v(x, t) \equiv x^\star P(t)x$ is bounded above and below by locally positive definite functions of x (criteria referred to as decrescence and positive definiteness, respectively.
- $-\dot{v}(x, t)$ is bounded below by some locally positive definite function. (Roughly speaking, $v(x, t)$ must be changing with respect to time at an increasing rate.)

The Basic Lyapunov Theorem implies that, if these conditions hold, then $x = 0$ is uniformly asymptotically stable, and is thus exponentially stable (by the equivalence of these two notions of stability for linear systems).

Fortunately, by hypothesis, we have:

$$\begin{aligned}
&\alpha I \leq P(t) \leq \beta I \\
\Rightarrow &\alpha |x|^2 \leq \underbrace{x^\star P(t) x}_{\equiv v(x,t)} \leq \beta |x|^2
\end{aligned}$$

where we have left multiplied each term by $x^\star$ and right multiplied each term by x to obtain the second expression.

In addition, we have:

$$\begin{aligned}\dot{v}(x,t) &= \dot{x}^\star P(t)x + x^\star \dot{P}(t)x + x^\star P(t)\dot{x} \\ &= x^\star(\dot{P}(t) + A^\star(t)P(t) + P(t)A(t))x \\ &= -x^\star Q(t)x \\ &\leq -\alpha|x|^2\end{aligned}$$

The proof is done.

"(1) $\Rightarrow$ (2)" : (see Appendix) This portion of the proof, which requires Lyapunov's Lemma, is placed in the Appendix.

■

The time-varying Lyapunov Lemma, as presented above, can be used to demonstrate the boundedness of $P(t)$ *if the given system is time-invariant.* However, we will directly demonstrate this below, since we presented an incomplete version of the proof for the time-varying Lyapunov Lemma.

Lemma 4.48. *Consider the system $\dot{x} = Ax$, where $A \in \mathbb{R}^{n\times n}$ and $\sigma(A) \in \mathbb{C}^-$. Then the unique solution to Lyapunov's Equation, $A^\star P + PA = -Q$, where $Q > 0$, is given by:*

$$P = \int_0^\infty e^{A^\star t} Q e^{At}\, dt$$

In particular, the above integral is well-defined.

Proof. Define the time-dependent matrix:

$$\begin{aligned}S(t) &\equiv \int_0^\star e^{A^\star \tau} Q e^{A\tau}\, d\tau \\ &= \int_0^\star e^{A^\star(t-\tau)} Q e^{A(t-\tau)}\, d\tau\end{aligned}$$

Differentiating $S(t)$ with respect to t (by again applying differentiation under the integral sign), we have:

$$\dot{S}(t) = A^\star S + SA + Q.$$

Since $\sigma(A) \in \mathbb{C}^-$, the terms $e^{A^\star(t-\tau)}$ and $e^{A(t-\tau)}$ are bounded above by exponential terms, so:

$$P \equiv S(\infty) = \int_0^\infty e^{A^\star \tau} Q e^{A\tau}\, d\tau$$

is well-defined. The existence (convergence) of P implies that:

$$A^\star P + PA + Q = \lim_{t\to\infty} \dot{S}(t) = 0.$$

It remains to demonstrate the uniqueness of P. Consider the linear mapping $L : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$, defined by:

$$L(X) = A^\star X + XA.$$

Let $M \in \mathbb{R}^{n\times n}$. Repeating the above procedure, we find that:

$$\tilde{X} \equiv \int_0^\infty e^{\tau A^\star}(-M)e^{\tau A}\,d\tau$$

satisfies $L(\tilde{X}) = M$. Thus, L is surjective. Since the domain and codomain of the linear mapping L have the same dimension, this implies that L is injective. This establishes the uniqueness of P.

■

Remark. See the appendix for a more direct, but also more mathematically intensive, proof of the uniqueness of P.

Theorem 4.49 (**Lyapunov Lemma** (Theorem 5.36, pgs. 211-212))**.** *Let $A, P, Q \in \mathbb{R}^{n\times n}$, with $Q > 0$, and consider the matrix equation:*

$$A^\star P + PA = -Q. \tag{4.4}$$

The following statements are equivalent:

1. *$\dot{x} = Ax$ is exponentially stable.*
2. *$\sigma(A) \subset \mathbb{C}^-$.*
3. *There exists some $Q > 0$ such that:*

   $$A^\star P + PA = -Q$$

   admits a unique solution $P > 0$.
4. *For each $Q > 0$:*

   $$A^\star P + PA = -Q$$

   admits a unique solution $P > 0$.

Proof. We have already established that (1) ⇔ (2), and it is trivial that (4) ⇒ (3). Lemma **??** establishes (3) ⇒ (1). It remains to show that (2) ⇒ (4).

The lemma above implies that, for linear, time-invariant systems, the solution is of the form:

$$P = \int_0^\infty e^{A^\star t}Qe^{At}\,dt$$

It is clear that $P \geq 0$. To show that $P > 0$, first note that since $Q > 0$, there exists some non-singular M such that $Q = M^\star M$. Now, let $x \neq 0$ be arbitrarily given such that:

$$0 = x^\star Px = \int_0^\infty x^\star e^{A^\star t}Qe^{At}x\,dt = \int_0^\infty |Me^{At}x|^2\,dt$$

Thus, $Me^{At}x$. Since M and e^{At} are non-singular, we have $x = 0$. This verifies that $P > 0$. ■

Alternative proof to Lyapunov's Theorem, "(4) $\Rightarrow$ (2)":

Proof. Consider the value function $V : \mathbb{R}^n \to \mathbb{R}$ given by:

$$V(x,t) = x^\star P x$$

Given $Q > 0$, let $P > 0$ be the solution to Lyapunov's Equation, $A^\star P + PA = -Q$. As derived above, we have:

$$\dot{V}(x,t) = -x^\star Q x.$$

We proceed to bound $\dot{V}(x,t)$ by $V(x,t)$ to establish an exponential limit for $V(x)$, and, in turn, $|x|$. Since, $P, Q > 0$:

$$\lambda_{\min}(P) \cdot |x|^2 \leq \underbrace{x^\star P x}_{\equiv V(x,t)} \leq \lambda_{\max}(P) \cdot |x|^2$$
$$\lambda_{\min}(Q) \cdot |x|^2 \leq \underbrace{x^\star Q x}_{\equiv \dot{V}(x,t)} \leq \lambda_{\max}(Q) \cdot |x|^2$$

We thus have:

$$\dot{V}(x) = -x^\star Q x \leq -\lambda_{\min}(Q) \cdot |x|^2 \leq -\underbrace{\frac{\lambda_{\min}(Q)}{\lambda_{\max}(P)}}_{\equiv k_1} \cdot V(x)$$
$$\Rightarrow V(x) \leq V(x_0) \cdot e^{-kt}$$

where we have defined $k_1 \equiv \lambda_{\min}(Q)/\lambda_{\max}(P)$. But $V(x) = x^\star P x$, so $V(x_0) = x_0^\star P x_0$, and thus the above equation implies:

$$\lambda_{\min}(P) \cdot |x|^2 \leq V(x) \leq V(x_0) \cdot e^{-kt} \leq \lambda_{\max}(P) \cdot |x_0|^2 e^{-kt}$$

Taking $k_2 = \lambda_{\max}(P)/\lambda_{\min}(P)$, we have:

$$|x(t)| \leq \sqrt{k_2} e^{-\frac{1}{2}k_1 t}$$

Thus, $\dot{x} = Ax$ is exponentially stable. ■

Numerical computation using Part 4 of the above lemma, which essentially involves solving an $n \times n$ system of linear equations, is more efficient than Part 2 of the above lemma, which involves finding the roots of an n-degree polynomial.

We conclude this section by stating without proof a generalization of the Lyapunov lemma, called the Taussky Lemma. It is useful when $\sigma(A) \not\subset \mathbb{C}^-$.

Theorem 4.50 (**Taussky Lemma** Lemma 5.37, pg. 212)**.** *Let $A, Q \in \mathbb{R}^{n\times n}$ such that $Q > 0$. If A has no eigenvalues on the imaginary axis, then the unique symmetric solution P to the Lyapunov Equation:*

$$A^\star P + PA = -Q$$

has as many positive eigenvalues as the number of eigenvalues of A in $\mathbb{C}^-$, and as many negative eigenvalues as the number of eigenvalues of A in $\mathbb{C}^+$.

Proof. See "O. Taussky. *A remark on a theorem of Lyapunov.* Journal of Mathematical Analysis and Applications. 2: 015-107, 1961. ■

Instability Theorem:

The Basic Lyapunov Theorem tells us that, if the value function $v(x,t)$ satisfies certain constraints, then the system satisfies certain notions of stability. A partial converse is stated in the following theorem, i.e. if the valued function $v(x,t)$ is "sufficiently ill-behaved" in a certain sense, the system is guaranteed to be unstable. This theorem will be used again to prove the instability of non-linear systems that can be approximated as unstable linear systems, via the indirect method of Lyapunov (see Appendix).

Theorem 4.51 (**Basic Instability Theorem** (Theorem 5.29, pg. 206))**.** *If there exists a value function $v(x,t)$ satisfying each of the conditions below:*

1. *$v(x,t)$ is decrescent,*
2. *$\dot{v}(x,t)$ is locally positive definite,*
3. *There exist points x arbitrarily close to 0 such that $v(x,t_0) > 0$,*

then the equilibrium point 0 is unstable at time t_0.

Proof. Since $v(x,t)$ is decrescent and $\dot{v}(x,t)$ is locally positive definite, there exist $r, s > 0$ and $\alpha, \beta \in K$ such that:

$$\begin{aligned} v(x,t) &\leq \beta(|x|), \qquad x \in B_r, \\ \dot{v}(x,t) &\geq \alpha(|x|), \qquad x \in B_s. \end{aligned}$$

To show that 0 is an unstable equilibrium point, we must prove there exists some $\epsilon > 0$ such that, for every $\delta > 0$, there exists some $x_0 \in B_\delta$ and some corresponding time $T_\delta \geq t_0$ such that if $|x_0| < \delta$, then $x(T_\delta) \geq \epsilon$.

Choose $\epsilon = \min\{r, s\}$, and fix $\delta > 0$ arbitrarily. By the third condition on $v(x,t)$, there exists some $x_0 \in B_\delta$ such that $v(x_0, t_0) > 0$. If $|x_0| \geq \epsilon$, the proof is completed by taking $T_\delta = t_0$. Otherwise, $x_0 \in B_\epsilon = B_r \cap B_s$. Now, suppose by contradiction that $x(t) \in B_\epsilon$ for each $t \geq t_0$. Then we have:

$$\begin{aligned} &\dot{v}(x(t),t) \geq 0, \\ \Rightarrow &v(x(t),t) \geq v(x_0,t_0) > 0. \end{aligned}$$

Since $v(x,t)$ is decrescent, and therefore continuous, there must exist some $\delta' > 0$ such that $|x(t)| < \delta'$ implies $v(x(t),t) < v(x_0,t_0)$. The above inequality states that $v(x(t),t) \geq v(x_0,t_0)$, so we must have $|x(t)| \geq \delta'$. Thus:

$$\begin{aligned} &\dot{v}(x,t) \geq \alpha(|x|) \geq \alpha(\delta') > 0, \\ \Rightarrow\ v(x(t),t)) &= v(x_0,t_0) + \int_{t_0}^{t} \dot{v}(x(\tau),\tau) d\tau \\ &\geq v(x_0,t_0) + \alpha(\delta') \cdot (t - t_0). \end{aligned}$$

In particular, when:

$$t = t_0 + \frac{\epsilon - v(x_0, t_0)}{\alpha(\delta')}$$

we have $v(x(t), t) \geq \epsilon$, a contradiction. The proof is done.

■

Indirect Lyapunov's Method:

If a system is only slightly non-linear, in the sense that there exists some $A(t)$ such that:

$$\dot{x} = A(t)x + f_1(x,t), \qquad \text{with :}$$
$$\lim_{|x|\to 0} \sup_{t\geq 0} \frac{|f_1(x,t)|}{|x|} = 0,$$

then *local* uniform asymptotic stability can be determined by examining the *global* uniform asymptotic stability of the linear system:

$$\dot{z} = A(t)z$$

Roughly speaking, we derive a slightly weaker sense of stability for a slightly non-linear system from the stability of the linear system that it resembles. This is rigorously justified by the following theorem, which also establishes a partial converse—If a slightly non-linear system can be approximated as an unstable linear system, then that slightly non-linear system must also be unstable.

Theorem 4.52 (**Indirect Lyapunov's Method** (Theorems 5.41, 5.42, pgs. 215-217))**.** *Suppose the non-linear system $\dot{x} = f(x,t)$, $x(t_0) = x_0$ has the linear approximation:*

$$\dot{x} = f(x,t) = A(t)x + f_1(x,t), \qquad \textit{with}$$
$$\lim_{|x|\to 0} \sup_{t\geq 0} \frac{|f_1(x,t)|}{|x|} = 0.$$

Then the following statements hold:

1. *If $\frac{\partial f(x,\cdot)}{\partial x}\big|_{x=0}$ is bounded in time, and 0 is a uniformly asymptotically stable equilibrium point of the linearized system:*

$$\dot{\hat{z}}(t) = \left.\frac{\partial f_1(x,t)}{\partial x}\right|_{x=0} z(t),$$

   then 0 is also a locally uniformly asymptotically stable equilibrium point of the original non-linear system $\dot{x} = f(x,t)$, $x(t_0) = x_0$.

2. *If $\frac{\partial f(x,\cdot)}{\partial x}\big|_{x=0}$ is constant in time, and has at least one eigenvalue in $\mathbb{C}^+$, then 0 is an unstable equilibrium point of the original nonlinear system $\dot{x} = f(x,t)$, $x(t_0) = x_0$.*

Proof. (See Appendix). ■

4.8 Lecture 15 Discussion

Example (**Discussion 10, Problem 2**). Show that, if $A(t) = -A(t)^T$, then $\dot{x}(t) = A(t)x(t)$ is internally stable.

Solution:

We directly evaluate the time derivative of the square of the norm of the state:

$$\begin{aligned}\frac{d}{dt}|x|^2 &= \dot{x}^T x + x^T \dot{x} = x^T A^T x + x^T A x \\ &= x^T (A^T + A)x = 0.\end{aligned}$$

Thus, $|x|$ is constant with respect to time, so the system is stable.

Example (**Discussion 10, pg. 6, Fall 2009 Prelims, Prof. Arcak**). Consider the LTI system:

$$\begin{aligned}\dot{x} &= Ax + Bu, \\ y &= Cx.\end{aligned}$$

1. Suppose there exists some $P > 0$ and a constant α such that:

$$A^T P + PA < \alpha P \tag{4.5}$$

   (a) Which region in the complex plane do the eigenvalues of A lie in?

   (b) Suppose (4.5) holds with $\alpha = 0$, and in addition, $PB = C^T$. Show that the given system is asymptotically stable for any feedback $u = -ky$, where $k \geq 0$.

2. Suppose, instead of (4.5), $P > 0$ satisfies the equality:

$$A^T P + PA = O$$

   (a) Which region in the complex plane do the eigenvalues of A lie in?

   (b) Does the above equality guarantee asymptotic stability for the feedback $u = -ky$, with a positive gain $k > 0$? If not, what additional conditions would you need?

Solution:

1. (a) Let $\lambda \in \sigma(A)$; then there exists some $v \neq 0$ such that $Av = \lambda v$. Substituting into (4.5), we have:

$$\begin{aligned}0 > v^\star (A^\star P + PA - \alpha P)v &= (2\mathrm{Re}(\lambda) - \alpha) v^\star P v \\ \Rightarrow \mathrm{Re}\lambda &< \frac{1}{2}\alpha.\end{aligned}$$

(b) Substituting $u = -ky = -kCx$ into the given system, we have:

$$\dot{x} = (A - kBC)x$$

Since (4.5) holds with $\alpha = 0$, we have:

$$\begin{aligned} & A^\star P + PA < 0, \\ \Rightarrow & (A - kBC)^\star P + P(A - kBC) + k(C^\star B^\star P + PBC) < 0, \\ \Rightarrow & (A - kBC)^\star P + P(A - kBC) + 2kC^\star C < 0, \\ \Rightarrow & (A - kBC)^\star P + P(A - kBC) < 0, \end{aligned}$$

since $C^\star C \geq 0$. The above result thus implies that, for each $\lambda \in \sigma(A - kBC)$, we have Re $\lambda < \frac{1}{2}\alpha = 0$. Thus, the given feedback renders the system asymptotic stable.

2. (a) Our solution for 1 a) implies that Re $\lambda = 0$ for each $\lambda \in \sigma(A)$. In other words, the eigenvalues of A lie on the imaginary axis.

   (b) Our solution for 1 b) implies that:

$$(A - kBC)^\star P + P(A - kBC) + 2kC^\star C = O$$

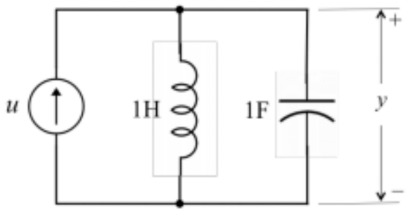
To achieve asymptotic stability, we require

$$(A - kBC)^\star P + P(A - kBC) < 0,$$

i.e. $C^\star C > 0$. This occurs if and only if C has full column rank.

Example (**Discussion 10, pg. 7, Fall 2015 Prelims, Prof. Carmena**).

1. Is the network shown in the figure BIBO stable? If not, find a bounded input that will excite an unbounded output.



2. Is the homogeneous state equation shown below asymptotically stable? Marginally stable?

$$\dot{x} = \begin{bmatrix} -1 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} x$$

3. Is the following state equation controllable? Observable? If not, reduce it to a controllable and observable form.

Solution :

1. Let i_1 and i_2 be the current flowing through the $1H$ inductor and the $1F$ capacitor, respectively, both in the downward direction. By Kirchhoff's Circuit Law (KCL) and Kirchhoff's Voltage Law (KVL), we have:

$$\begin{aligned} y &= L\frac{di_1}{dt} = \frac{1}{C}\int_0^t i_2\,dt, \\ u &= i_1 + i_2 \end{aligned}$$

Define the states of the systems to be $x_1 = i_1, x_2 = \dot{i}_1$. Then:

$$\begin{aligned} \dot{x}_1 &= x_2, \\ \dot{x}_2 &= \frac{1}{L}\dot{y} = \frac{1}{LC}(u - i_1) = -\frac{1}{LC}x_1 + \frac{1}{LC}u, \\ y &= L\dot{i}_1 = Lx_2. \end{aligned}$$

In matrix form, we have:

$$\begin{aligned} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} &= \underbrace{\begin{bmatrix} 0 & 1 \\ -\frac{1}{LC} & 0 \end{bmatrix}}_{\equiv A}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \underbrace{\begin{bmatrix} 0 \\ \frac{1}{LC} \end{bmatrix}}_{\equiv B} u \\ y &= \underbrace{\begin{bmatrix} 0 & L \end{bmatrix}}_{\equiv C}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \end{aligned}$$

The transfer function of the system is:

$$\begin{aligned} H(s) &= B(sI - A)^{-1}C \\ &= \begin{bmatrix} 0 & L \end{bmatrix}\begin{bmatrix} s & -1 \\ \frac{1}{LC} & s \end{bmatrix}^{-1}\begin{bmatrix} 0 \\ \frac{1}{LC} \end{bmatrix} \\ &= \frac{1}{C}\cdot\frac{s}{s^2 + \frac{1}{LC}}, \end{aligned}$$

with poles at $\pm i1/\sqrt{LC}$, both of which are on the imaginary axis. Thus, the system is not BIBO stable. Since its poles are at $\pm i1/\sqrt{LC}$, an example of a bounded input that would excite an unbounded output would be a sinuosidal function with angular frequency $\frac{1}{\sqrt{LC}}$.

2. Observe that:

$$A \equiv \begin{bmatrix} -1 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Then $\dot{x} = Ax$.

Since A is upper triangular, its eigenvalues can be read off the diagonal—$\sigma(A) = \{-1, 0\}$. The eigenvalue 0 is repeated twice, but each is associated with a Jordan block of size 1. Thus, the system is stable (although not asymptotically stable).

3. By inspection, when $s = \lambda_1$, the third row of the controllability matrix pencil $\begin{bmatrix} sI - A & B \end{bmatrix}$ is a zero row, and the first column of $\begin{bmatrix} sI - A \\ C \end{bmatrix}$ is a zero column. Thus, $\begin{bmatrix} sI - A & B \end{bmatrix}$ and $\begin{bmatrix} sI - A \\ C \end{bmatrix}$ lack full row rank and full column rank, respectively, when $s = \lambda_1$; as a result, the given state equation is neither controllable nor observable.

Example (**Discussion 10, pg. 8, Spring 2017 Prelims, Prof. El Ghaoui**). Consider a continuous-time LTI system $\dot{x}(t) = Ax(t), t \geq 0$, with no input (such a system is said to be autonomous), and output $y(t) = Cx$. We wish to evaluate the energy contained in the system's output, as measured by the index:

$$J(x_0) \equiv \int_0^\infty y(t)^T y(t)\, dt = \int_0^\infty x(t)^T Q x(t)\, dt$$

where $Q \equiv C^T C \succeq 0$.

1. Show that if the system is stable, then $J(x_0) < \infty$ for any given x_0. *Hint*: Show that $\|y(t)\|_2 \leq c\|x_0\|_2 e^{\sigma_{\max} t}$, where $\sigma_{\max}$ is the maximum real part of the eigenvalues λ_i of A, and $c > 0$ is some constant.

2. Show that if the system is stable and there exists a matrix $P \succeq 0$ such that:

$$A^T P + PA + Q \preceq 0,$$

then it holds that $J(x_0) \leq x_0^T P x_0$. *Hint*: Consider the quadratic form $V(x(t)) = x(t)^T P x(t)$, and evaluate its derivative with respect to time.

3. Explain how to compute a minimal upper bound on the state energy, for the given initial conditions.

Remark. Here, we interpret "stable" as "asymptotically stable."

Solutions :

We also assume that $A \in \mathbb{R}^{n \times n}$, $C \in \mathbb{R}^{p \times n}$, and that, in general, $p \neq n$.

1. Observe that:

$$\begin{aligned} y(t) = Cx(t) = Ce^{tA}x_0, \\ \Rightarrow J(x_0) = \int_0^\infty y(t)^T y(t)\, dt = \int_0^\infty \|Ce^{tA}x_0\|^2 dt \end{aligned}$$

Now, since the system is stable, there exists some $\sigma_{\max} > 0$ such that all eigenvalues of e^{tA} decay exponentially at a rate faster than $\sigma_{\max}$. Thus, since each term in $Ce^{tA}x_0$ is a linear combination of $e^{\lambda_i t}$, where $\sigma(A) = \{\lambda_i | i = 1, \cdots, n\}$, it follows that the infinite integral $J(x_0)$ converges.

2. Following the hint, we have:

$$\begin{aligned} &\because \frac{d}{dt}(x^T P x) = \dot{x}^T P x + x^T P \dot{x} = x^T (A^T P + PA) x \leq -x^T Q x, \\ &\Rightarrow J(x_0) = \int_0^\infty x^T Q x \, dt \leq \int_0^\infty -\frac{d}{dt}(x^T P x) \, dt = x_0^T P x_0 \end{aligned}$$

3. Let $\overline{x_0} = Px_0$, where P is an invertible matrix whose columns are the eigenvectors (or generalized eigenvectors) of A, in a correct order i.e. $A = P^{-1}JP$ for some square matrix J in Jordan form. For convenience, we choose to focus on the 1-norm of the state energy, defined by:

$$\begin{aligned} \int_0^\infty |x(t)|_1^2 \, dt &= \int_0^\infty \left| e^{tA} x_0 \right|_1^2 dt = \int_0^\infty |P^{-1} e^{tJ} P x_0|_1^2 \, dt \\ &\leq \|P^{-1}\|_1 \cdot \int_0^\infty \|e^{tJ}\|_1^2 \, dt \cdot |P x_0|_1^2 \\ &\leq \|P^{-1}\|_1 \cdot \int_0^\infty \left(\sum_{k=0}^{n-1} \frac{\sigma_{\max}^k t^k}{k!} \right)^2 e^{-2\sigma_{\max} t} \, dt \cdot |P x_0|_1^2 \end{aligned}$$

where, in the worst case, J may contain a single Jordan block of size n, and the maximum column sum of e^{tJ} thus contains n non-zero elements of the form $\frac{\sigma_{\max}^k t^k}{k!}$, for each $k = 0, 1, \cdots, n-1$.

Example (**Discussion 10, pg. 9, Fall 2013 Prelims, Prof. Arcak**).

1. Consider the linear system $\dot{x} = Ax$, where:

$$A = \begin{bmatrix} \alpha & \beta \\ -\beta & \alpha \end{bmatrix},$$

with $\alpha, \beta \in \mathbb{R}$. For which values of α, β is the system stable, asymptotically stable, and unstable?

2. Suppose $\alpha = 0$. If we fix some $T > 0$ and take samples of the trajectories every T units of time, we obtain:

$$x[n] \equiv x(nT)$$

for each $n = 0, 1, 2, \cdots$.

(a) Find the matrix A_d for the discrete-time model $x[n+1] = A_d x[n]$ as a function of β and T.

(b) For which values of β and T are the solutions $x[n]$ periodic in n?

Solution:

1. The characteristic function of A is:

$$\chi_A(s) \equiv \det(sI - A) = (s-\alpha)^2 + \beta^2$$

We thus have $\sigma(A) = \{\alpha \pm \beta i\}$. If $\alpha < 0$, the system is asymptotically stable; if $\alpha > 0$, the system is unstable. If $\alpha = 0$, and $\beta \neq 0$, then the system state oscillates sinusoidally; if $\alpha = \beta = 0$, then $A = 0$, so $x(t)$ remains constant. In summary, the system is stable, asymptotically stable, and unstable when $\alpha = 0$, $\alpha < 0$, and $\alpha > 0$, respectively.

2. Since the original system is $\dot{x} = Ax$, we have $x(t) = e^{(t-t_0)A}x(t_0)$.

(a) Taking $t_0 = nT$ and $t = (n+1)T$, we have:

$$x[n+1] = x((n+1)T) = e^{TA}x[n]$$

Thus, $A_d = e^{TA}$. We can evaluate this matrix exponential using the Cayley-Hamilton Theorem; notice that, with $\alpha = 0$, the characteristic equation becomes $\chi_A(s) = s^2 + \beta^2$. Let $q(s)$ and $a_1, a_0 \in \mathbb{C}$ be given such that:

$$e^{Ts} = (s^2 + \beta^2) \cdot q(s) + \alpha_1 s + \alpha_0,$$

Substituting $s = \pm i\beta$, we have:

$$e^{i\beta T} = i\alpha_1\beta + \alpha_0$$
$$e^{-i\beta T} = -i\alpha_1\beta + \alpha_0$$

We thus have $\alpha_1 = \frac{1}{\beta}\sin\beta T$ and $\alpha_2 = \cos\beta T$, so

$$e^{TA} = \left(\frac{1}{\beta}\sin\beta T\right)A + (\cos\beta T)\,I = \begin{bmatrix} \cos\beta T & \frac{1}{\beta}\sin\beta T \\ \frac{1}{\beta}\sin\beta T & \cos\beta T \end{bmatrix}$$

(b) If $x[n]$ is periodic in n, there must exist some positive integer $N \in \mathbb{N}$ for which $x[n+N] = x[n]$ for each $n = 0, 1, 2, \cdots$. Since $x[n+N] = e^{NTA} \cdot x[n]$, we must have $e^{NTA} = I$.

Repeating the above procedure, we find that e^{NTA} can be found by simply replacing T with NT in the expression for e^{TA}, i.e.:

$$e^{NTA} = \begin{bmatrix} \cos N\beta T & \frac{1}{\beta}\sin N\beta T \\ \frac{1}{\beta}\sin N\beta T & \cos N\beta T \end{bmatrix}$$

Setting $e^{N\beta T} = I$, we find that $\beta T = \frac{2\pi}{N}$. In other words, $x[n]$ is periodic in n if and only if βT is a positive rational multiple of 2π.

Example (**Discussion 10, pg. 10, Spring 2017 Prelims, Prof. Fearing**).

1. Given an LTI system $\dot{x} = Ax$, and a matrix $M > 0$ such that $V \equiv x^T M x$ satisfies $\dot{V} < 0$ for any trajectory, determine the possible range of eigenvalues for A.

2. Consider the continuous-time linear system defined by:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1 \\ 2 \end{bmatrix} u$$

   (a) For $u(t) = 0$, determine the state trajectory with:

$$x_a(t=0) = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \qquad x_b(t=0) = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

   (b) For $u(t) = 1$ for $t \geq 0$, determine the state trajectory with:

$$x_a(t=0) = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \qquad x_b(t=0) = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

   (c) Given an initial condition x_0, explain how you would find a $u(t)$ such that $x(t)$ asymptotically approaches a finite fixed value, say for:

$$x_0 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \qquad x_f = \begin{bmatrix} 2 \\ 0 \end{bmatrix}.$$

   (d) Given an initial condition x_0, explain whether it is possible to find a $u(t)$ such that $x(t)$ asymptotically approaches a finite fixed value, say for:

$$x_0 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \qquad x_f = \begin{bmatrix} 2 \\ 0 \end{bmatrix},$$

   with fixed $x_1(t) = 2$ for $t \geq 0$?

Solution:

1. Differentiating V with respect to time, we have:

$$0 > \dot{V} = \dot{x}^T M x + x^T M \dot{x} = x^T (AM + MA) x,$$

   where $x \in \mathbb{R}^n$ is arbitrary. Let $\lambda \in \sigma(A)$, and let v be a corresponding eigenvector. Then:

$$0 > v^T (AM + MA) v = v^T (2\lambda M) v.$$

   Since $v^T M v > 0$, we have $\lambda < 0$.

2. (a) If $u = 0$, then:

$$\dot{x}_1 = -x_1,$$
$$\dot{x}_2 = -2x_2,$$

Thus, we have:

$$x_1(t) = x_1(0)e^{-t}$$
$$x_2(t) = x_2(0)e^{-2t}$$

(b) If $u = 1$, then:

$$\dot{x}_1 = -x_1 + 1,$$
$$\dot{x}_2 = -2x_2 + 2,$$

Thus, we have:

$$x_1(t) = x_1(0)e^{-t} + t$$
$$x_2(t) = x_2(0)e^{-2t} + 2t$$

(c) Suppose by contradiction that there exists a function $u(\cdot)$ that drives the system from $x_0 = \begin{bmatrix} 2 & 1 \end{bmatrix}^T$ to $x_f = \begin{bmatrix} 2 & 0 \end{bmatrix}^T$ Then:

$$\dot{x}_1 = -x_1 + u(t),$$
$$\dot{x}_2 = -2x_2 + 2u(t),$$

Rearranging terms and taking $t \to \infty$, the above two equations become contradictory:

$$\lim_{t\to\infty} u(t) = \lim_{t\to\infty} \dot{x}_1(t) + \lim_{t\to\infty} x_1(t) = 0 + 2 = 2,$$
$$\lim_{t\to\infty} u(t) = \frac{1}{2} \cdot \lim_{t\to\infty} \dot{x}_2(t) + \lim_{t\to\infty} x_2(t) = 0 + 0 = 0.$$

Thus, there exists no input driving $x(t)$ from x_0 to x_f asymptotically.

(d) The answer to c implies that there exist no such $u(\cdot)$. However, even without the answer to c, we can show this to be true.
Suppose by contradiction that such an input u exists. Since:

$$\dot{x}_1 = -x_1 + u,$$

and we want $x_1(t) = 2, \dot{x}_1(t) = 0$ at each time t, we must apply $u = 2$, with the result that:

$$\dot{x}_2 = -2x_2 + 4$$

Since $x_2(0) = 1$, we thus have:

$$x_2(t) = e^{-2t} + 4t$$

as the unique solution to $x_2(t)$. Thus, $x_2(t) \to \infty$ as $t \to \infty$, contradicting the fact that $x(t)$ is supposed to asymptotically approach $x_f = \begin{bmatrix} 2 & 0 \end{bmatrix}^T$.

Chapter 5

Controllability and Observability

5.1 Lecture 16

In this lecture, we wish to understand how, given a dynamical system, we can design an input $u(t)$ such that the system produces a desirable output $y(t)$. Notice that $u(t)$ affects $y(t)$ through $x(t)$.

Note. Below, u_τ will denote the function u in the time interval τ. Notice that τ is often an open, closed, or half-open-half-closed interval, e.g. (t_0, t), $[t_0, t]$, or $[t_0, t)$.

Definition 5.1 (**Steering**). *Let $(U, \Sigma, \mathcal{Y}, s, r)$ be a dynamical system representation, and let t_0, t_1 be given with $t_0 < t_1$. The input $u_{[t_0,t_1]}(\cdot)$ **steers** (x_0, t_0) **to** (x_1, t_1) if:*

$$x_1 = s(t_1, t_0, x_0, u_{[t_0,t_1]})$$

Definition 5.2 (**(Complete) Controllability on** $[t_0, t_1]$). *The system representation D is **(completely) controllable on** $[t_0, t_1]$ if, for each $x_0, x_1 \in \Sigma$, there exists some $u_{[t_0,t_1]} \in \mathcal{U}$ that steers x_0 at t_0 to x_1 at t_1.*

Proposition 5.3. *Given a dynamical system $D = (U, \Sigma, \mathcal{Y}, s, r)$, the following are equivalent:*

1. *D is controllable on $[t_0, t_1]$.*
2. *For each $x_0 \in \Sigma$, the map:*

$$x(\cdot) = s(t_1, t_0, x_0, u_{[t_0,t_1]}(\cdot)) : U \to \Sigma$$

is surjective.

Memoryless Feedback and Controllability:

In general, the input $u(\cdot)$ may be an output feedback, i.e. $u(y, t)$, or a state feedback, i.e. $u(x, t)$. Consider two memoryless maps:

$$F_S : \Sigma \to \mathcal{U}$$
$$F_O : \mathcal{Y} \to \mathcal{U}$$

and the following block diagrams, which represent state feedback and output feedback, respectively:

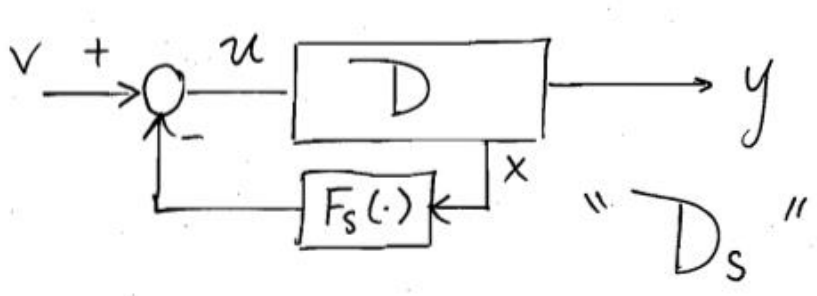


Figure 5.1: State Feedback

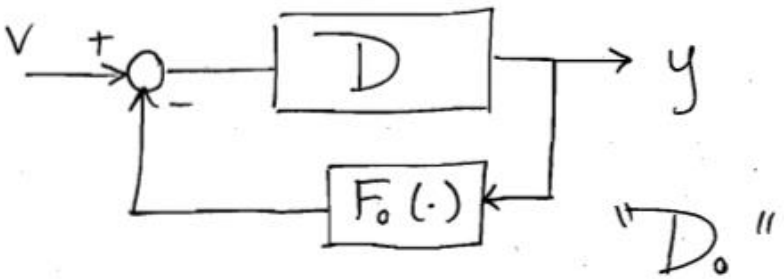


Figure 5.2: Output Feedback

In the top and bottom figures, the input $u(\cdot)$ corresponds to the state and output feedback, respectively, whereas $v(\cdot)$ is known as the *auxilliary input*. This simply means that $v(\cdot)$ is the input we apply to the closed-loop system:

$$u(x,t) = v(t) - F_S(x(t))$$
$$u(y,t) = v(t) - F_O(y(t))$$

Below, we introduce several equivalent definitions for the controllability of a state.

We first discuss an important concept, known as the *well-posedness assumption*, on which the validity of many of these results depend.

Definition 5.4 (Well-Posed). *Let $D = (\mathcal{U}, \Sigma, \mathcal{Y}, s, r)$ be a dynamical system representation, and let D_S and D_O be memoryless feedback systems constructed from D, as given above. D is said to be **well-posed** if, for each initial state x_0, initial time t_0, and overall input $v(\cdot)$, the systems D_S and D_O have unique inputs and outputs.*

*If a system is not well-posed, it is said to be **ill-posed**.*

The well-posedness assumption captures the intuitive notion that, for the closed feedback loop to make sense, there should be some delay about the feedback loop. For example, if there were no delay in the closed state feedback loop, the input u at some time t would produce a state x, at the same time t, and then this state would be fed back to affect u, again at the same time t. This results in circular (and thus contradictory) logic.

Example (**An Ill-Posed System**). Suppose by contradiction that the following linear dynamical system:

$$\dot{x} = 0,$$
$$y = u,$$

with $F_s(\cdot) = F_o(\cdot) = -1$, is well-posed. Then $y = v + u$, a contradiction to the fact that $y = u$, and v is arbitrary.

Theorem 5.5. *Let $D = (\mathcal{U}, \Sigma, \mathcal{Y}, s, r)$ be a dynamical system representation, and let D_S and D_O be well-posed memoryless feedback systems constructed from D. Then the following statements are equivalent:*

1. *D is controllable on $[t_0, t_1]$.*
2. *D_S is controllable on $[t_0, t_1]$.*
3. *D_O is controllable on $[t_0, t_1]$.*

Proof.
"(1) $\Leftrightarrow$ (2)" :

We will first establish that (1) $\Rightarrow$ (2). Fix $x_0, x_1 \in \Sigma$. Since D is controllable on $[t_0, t_1]$, there exists some $\tilde{u}_{[t_0,t_1]}(\cdot)$ that steers x_0 at t_0 to x_1 at t_1. To demonstrate the controllability of D_S, we must show that there exists some input $\tilde{v}(t)$ to D_S that steers (x_0, t_0) to (x_1, t_1). This can be done by defining:

$$\begin{aligned}\tilde{v}(t) &= \tilde{u}(t) + F_S(x(t)) \\ &= \tilde{u}(t) + F_S(s(t, t_0, x_0, \tilde{u}_{[t_0,t]}))\end{aligned}$$

Conversely, if D_S were controllable, then for each $x_0, x_1 \in \Sigma$, there exists some control $\tilde{v}_{[t_0,t_1]}$ that steers (x_0, t_0) to (x_1, t_1) on D_S. Thus, the input $\tilde{u}(t)$ generated internally by D_S steers (x_0, t_0) to (x_1, t_1) on D_S, so D is controllable.

"(1) $\Leftrightarrow$ (3)" :

The proof here is similar to the proof for "(1) $\Leftrightarrow$ (2)"; simply replace the state feedback F_S with the output feedback F_O, and the state transition map $s(t, t_0, x_0, \tilde{u}_{[t_0,t_1]})$ with the response map $\rho(t, t_0, x_0, \tilde{u}_{[t_0,t_1]})$.

■

Now consider the case where the state feedback, instead of being a function mapping x to some control signal $F_S(x(t))$, is in fact a dynamical system F_S with representation:

$$F_S = (\Sigma, \Sigma_1, \mathcal{U}, s_F, r_F)$$

Contrast this with the original system $D = (\mathcal{U}, \Sigma, \mathcal{Y}, s, r)$. In short, F_S takes states $x \in \Sigma$ of D as inputs, and produce inputs $u \in \mathcal{U}$ of D as outputs. We denote F_S's own state space by Σ_1.

In fact, the above theorem holds even when the feedback loop is itself a dynamical system.

Theorem 5.6. *Let $D = (\mathcal{U}, \Sigma, \mathcal{Y}, s, r)$ be a dynamical system representation, and let D_S be a dynamical feedback system constructed from D. Then the following statements are equivalent:*

1. *D is controllable on $[t_0, t_1]$.*

2. *D_S is controllable on $[t_0, t_1]$.*

Definition 5.7 ((Complete) Observability on $[t_0, t_1]$). *The dynamical system $D = (U, \Sigma, \mathcal{Y}, s, r)$ is said to be **(completely) observable on** $[t_0, t_1]$ if, for each $u_{[t_0,t_1]}(\cdot) \in \mathcal{U}$ and each $y_{[t_0,t_1]}(\cdot) \in \mathcal{Y}$, the initial state $x_0 \equiv x(t_0)$ is uniquely determined by $u(\cdot)$ and $y(\cdot)$.*

Proposition 5.8. *The dynamical system $D = (U, \Sigma, \mathcal{Y}, s, r)$ is observable on $[t_0, t_1]$ if and only if the response map:*

$$y(\cdot) = \rho(\cdot, t_0, x_0, u_{[t_0,t_1]}(\cdot)) : \Sigma \to \mathcal{Y}$$

is injective.

Intuitively, this means that, given any output in a given time interval $[t_0, t_1]$, the state can be uniquely determined.

Memoryless Feedback and Memoryless Feedforward

Just as the output y can be connected to the input u via the feedback loop $F_o(\cdot)$, the input u can be "fed forward" to the output y via a *feedforward loop*, as shown below.

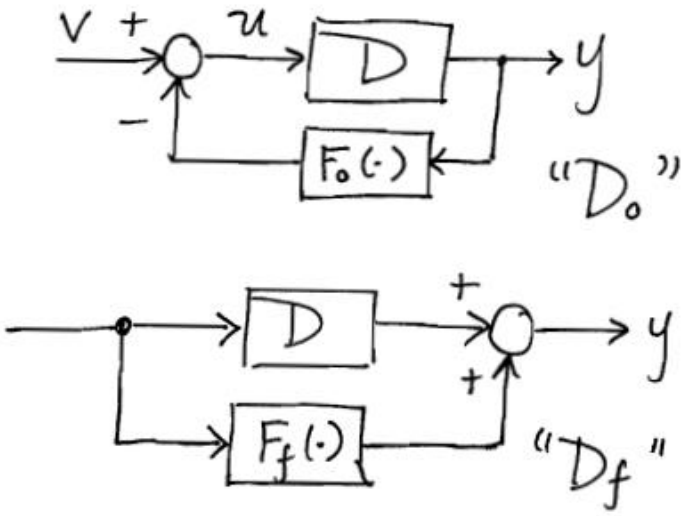


Figure 5.3: Memoryless Output Feedback and Memoryless Feedforward

It is a curious fact that neither memoryless output feedback nor memoryless feedfoward alters the observability of the system.

Theorem 5.9. *Let $D = (\mathcal{U}, \Sigma, \mathcal{Y}, s, r)$ be a dynamical system representation, and let D_o and D_f be well-posed memoryless output feedback and input feedforward systems constructed from D. Then the following statements are equivalent:*

1. *D_o is observable on $[t_0, t_1]$.*
2. *D_f is observable on $[t_0, t_1]$.*

However, *state feedback may affect observability.* For instance, it is possible to transform a completely observable system D into a closed-loop state feedback system D_S that is not completely observable. Here, D and D_S are defined as given above:

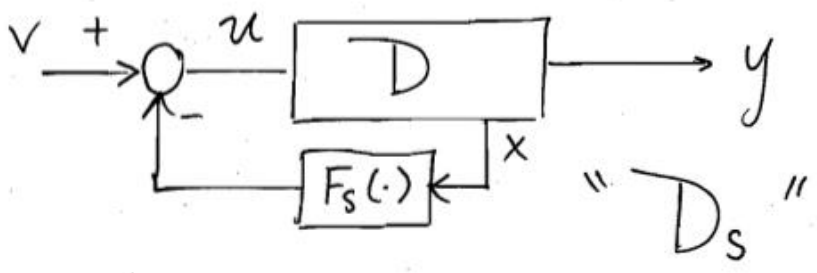


Figure 5.4: State Feedback

The reason is that the connection between the auxiliary input $v(\cdot)$ (i.e. the input to the closed-loop state feedback system D_S) and the actual input $u(\cdot)$ (i.e. the input into the original system D) is related via the system state x, which remains unknown to the observer.

Consider the counterexample below, which makes use of a state feedback chosen to place the state into the null space of the output matrix C.

Example (**State Feedback Changes Observability**). Consider the time-invariant system:

$$\begin{aligned}\dot{x} &= \underbrace{\begin{bmatrix}0 & 0\\ 1 & 0\end{bmatrix}}_{\equiv A} x + \underbrace{\begin{bmatrix}2\\ 1\end{bmatrix}}_{\equiv B} u,\\ y &= \underbrace{\begin{bmatrix}0 & 1\end{bmatrix}}_{\equiv C} x,\\ u &= \underbrace{\begin{bmatrix}1 & 0\end{bmatrix}}_{\equiv F_s} x\end{aligned}$$

Since the auxiliary input $v(\cdot)$ is given by:

$$v = u - F_s x$$

the dynamics of D can be rewritten as:

$$\begin{aligned}\dot{x} &= \begin{bmatrix}0 & 0\\ 1 & 0\end{bmatrix} x + \begin{bmatrix}2\\ 1\end{bmatrix}\left(v - \begin{bmatrix}1 & 0\end{bmatrix} x\right)\\ &= \underbrace{\begin{bmatrix}-2 & 0\\ 0 & 0\end{bmatrix}}_{\equiv A - BF_s} x + \underbrace{\begin{bmatrix}2\\ 1\end{bmatrix}}_{\equiv B} v,\\ y &= \begin{bmatrix}0 & 1\end{bmatrix} x,\end{aligned}$$

Using observability tests described in the next lecture, we can show that the original system D is completely observable, but the closed-loop system D_S is not.

5.2 Lecture 17

For a time-varying system $R = [A(\cdot), B(\cdot), C(\cdot), D(\cdot)]$, recall that the state transition and response maps as:

$$x(t) = \Phi(t, t_0)x_0 + \int_{t_0}^{t_1} \Phi(t, \tau)\, B(\tau)u(\tau)\, d\tau$$
$$y(t) = C(t)\Phi(t, t_0)x_0 + \int_{t_0}^{t_1} C(t)\Phi(t, \tau)\, B(\tau)u(\tau)\, d\tau + D(t)u(t)$$

Define the mappings:

$$L_c : \mathcal{U}_{[t_0,t_1]} \to \mathbb{R}^n$$
$$L_o : \mathbb{R}^n \to \mathcal{Y}_{[t_0,t_1]}$$

as follows. For each $t \geq 0$, we have:

$$L_c(u_{[t_0,t_1]}) = \int_{t_0}^{t_1} \Phi(t, \tau)B(\tau)u(\tau)\, d\tau$$
$$(L_o x_0)(\cdot) = C(\cdot)\Phi(\cdot, t_0)x_0$$
$$= y(\cdot) - \int_{t_0}^{\cdot} C(\cdot)\Phi(\cdot, t_0)B(\tau)u(\tau)\, d\tau - D(t)u(t)$$

Intuitively, L_c captures the notion that inputs $u(t)$ can be chosen to alter ("control") the state $x(t)$, whereas $y(t)$ can be obtained ("observed") through the state $x(t)$. In mathematical terms, $R(L_c)$ is the subspace of all states that can be controlled, while $N(L_o)$ is the subspace of all states that cannot be observed. This intuition suggests that $R(L_c)$ and $N(L_o)$ are related to the controllability and observability of the system, respectively. In particular, if L_c is surjective, and L_o is injective (i.e. $R(L_c) = \mathbb{R}^n$ and $N(L_o) = \{0\}$), then the system is completely controllable and completely observable, respectively. However, from a practical point of view, the surjectiveness of L_c and injectiveness of L_o are difficult to verify, since $R(L_c)$ and $N(L_o)$ are infinite-dimensional spaces.

Fortunately, from linear algebra, we know that:

$$R(L_c) = R(L_c L_c^\star)$$
$$N(L_o) = N(L_o^\star L_o)$$

Thus, instead of evaluating the dimensions of $R(L_c)$ and $N(L_o)$, we can instead attempt to evaluate $R(L_c L_c^\star)$ and $N(L_o^\star L_o)$. This is easier, since $L_c L_c^\star$ and $L_o^\star L_o$ are simply $n \times n$ (semi-positive definite) matrices. We call $W_c \equiv L_c L_c^\star$ and $W_0 \equiv L_o^\star L_o$ the *Controllability Grammian* and the *Observability Grammian*, respectively.

To evaluate W_c and W_o, we need to find suitable expressions for $L_c^\star$ and $L_o^\star$. These can be found using the original definition of the Hermitian adjoint of a vector. Let H_u denote the Hilbert space inhabited by inputs in the range $[t_0, t_1]$, and let $u_{[t_0,t_1]} \in H_u$ and $v \in \mathbb{R}^n$ be arbitrarily given. Then:

$$\begin{aligned}
\langle L_c^\star v, u_{[t_0,t]}\rangle_{H_u} &= \langle v, L_c u_{[t_0,t]}\rangle_{\mathbb{R}^n} \\
&= v^\star \int_{t_0}^{t} \Phi(t,\tau)B(\tau)u(\tau)\,d\tau \\
&= \int_{t_0}^{t} (B(\tau)^\star \Phi(t,\tau)^\star v)^\star u(\tau)\,d\tau \\
&= \left\langle B(\cdot)^\star \Phi(t,\cdot)^\star v, u_{[t_0,t]}\right\rangle_H \\
\Rightarrow\ (L_c^\star v)(\cdot) &= B(\cdot)^\star \Phi(t,\cdot)^\star v
\end{aligned}$$

Similarly, let H_y denote the Hilbert space inhabited by inputs in the range $[t_0, t_1]$, and let $y_{[t_0,t_1]} \in H_y$ and $v \in \mathbb{R}^n$ be arbitrarily given. Then:

$$\begin{aligned}
\langle L_o^\star y_{[t_0,t_1]}, v\rangle_{\mathbb{R}^n} &= \langle y_{[t_0,t_1]}, L_o v\rangle_{H_y} \\
&= \int_{t_0}^{t} y(\tau)^\star (L_o v)(\tau) \\
&= \int_{t_0}^{t} y(\tau)^\star C(\tau)\Phi(\tau,t_0)\,v\,d\tau \\
&= \left(\int_{t_0}^{t_1} \Phi(\tau,t_0)^\star C(\tau)^\star y(\tau)\right)^\star v \\
&= \left\langle \int_{t_0}^{t_1} \Phi(\tau,t_0)^\star C(\tau)^\star y(\tau)\,d\tau, v\right\rangle_{\mathbb{R}^n} \\
\Rightarrow\ L_0^\star y_{[t_0,t_1]} &= \int_{t_0}^{t_1} \Phi(\tau,t_0)^\star C(\tau)^\star y(\tau)\,d\tau
\end{aligned}$$

We thus have the following definitions.

Definition 5.10 (**Controllability Grammian, Observability Grammian**). *For a time-varying system $R = [A(\cdot), B(\cdot), C(\cdot), D(\cdot)]$, recall that the state transition and response maps as:*

$$\begin{aligned}
x(t) &= \Phi(t,t_0)x_0 + \int_{t_0}^{t} \Phi(t,\tau)\,B(\tau)u(\tau)\,d\tau \\
y(t) &= C(t)\Phi(t,t_0)x_0 + \int_{t_0}^{t} C(t)\Phi(t,\tau)\,B(\tau)u(\tau)\,d\tau + D(t)u(t)
\end{aligned}$$

*Define the **Controllability Grammian**, denoted by $W_c(t_0,t_1) : \mathbb{R}^n \to \mathbb{R}^n$ and the **Observability Grammian**, denoted by $W_o(t_0,t_1) : \mathbb{R}^n \to \mathbb{R}^n$ as:*

$$\begin{aligned}
W_c(t_0,t) \equiv L_c L_c^\star &= \int_{t_0}^{t} \Phi(t,\tau)\,B(\tau)\,B^\star(\tau)\,\Phi^\star(t,\tau)\,d\tau \\
W_o(t_0,t) \equiv L_o^\star L_o &= \int_{t_0}^{t} \Phi^\star(\tau,t_0)\,C^\star(\tau)\,C(\tau)\,\Phi(\tau,t_0)\,d\tau
\end{aligned}$$

respectively.

Remark. By definition, $W_c(t_0, t_1) = L_c L_c^\star$ and $W_o(t_0, t_1) = L_o L_o^\star$, so $W_c(t_0, t_1)$ and $W_o(t_0, t_1)$ are both semi-positive definite.

Theorem 5.11 (**Controllability of Linear Time-Variant Systems**). *For a time-varying system $R = [A(\cdot), B(\cdot), C(\cdot), D(\cdot)]$, the following statements are equivalent:*

1. *R is completely controllable (c.c.) on $[t_0, t_1]$.*
2. *For each $x_0 \in \mathbb{R}^n$, there exists some input $u_{[t_0,t_1]}$ that steers (x_0, t_0) to $(0, t_1)$.*
3. *For each $x_1 \in \mathbb{R}^n$, there exists some input $u_{[t_0,t_1]}$ that steers $(0, t_0)$ to (x_1, t_1).*
4. *The mapping $L_c : \mathcal{U}_{[t_0,t_1]} \to \mathbb{R}^n$ is surjective, i.e.:*
$$R(L_c) = \mathbb{R}^n$$
5. *The Controllability Grammian $W_c(t_0, t_1)$ is invertible for each t, i.e.:*
$$\operatorname{rank}(W_c(t_0, t_1)) = n$$
for each t. In fact, $W_c(t_0, t_1) > 0$.

Proof.

Fix t_1, t_0 such that $t_1 \geq t_0 \geq 0$, and arbitrarily $x_0, x_1 \in \Sigma$. Note that, if $x_1 = x(t)$, then:

$$x_1 = \Phi(t, t_0)x_0 + \int_{t_0}^{t} \Phi(t, \tau)B(\tau)u(\tau)\, d\tau \tag{5.1}$$

(1) $\Leftrightarrow$ (2) $\Leftrightarrow$ (3):

Note that the following are equivalent:

$$\begin{aligned} x_1 &= \Phi(t, t_0)x_0 + L_c u(\cdot) \\ 0 &= \Phi(t, t_0)\big(x_0 - \Phi^{-1}(t, t_0)x_0\big) + L_c u(\cdot) \\ x_1 - \Phi(t, t_0)x_0 &= L_c u(\cdot) \end{aligned}$$

Since x_0, x_1 can be any state in Σ, this demonstrates the equivalence of 1, 2, and 3.

(1) $\Leftrightarrow$ (4):

Since the state $x_0, x_1 \in \Sigma$ in (5.1) are arbitrarily chosen, a necessary condition for complete controllability is the surjectivity of $R(L_c)$:

$$R(L_c) = \mathbb{R}^n$$

Intuitively, this is required for complete controllability, since $x_0, x_1 \in \Sigma$ can be any vector in $\mathbb{R}^n$.

However, this is also sufficient, since if x_1 and x_0 are given at t_1 and t_0, respectively, then there exists some $y_{[t_0,t_1]}$ such that:

$$x_1 - \Phi(t, t_0)x_0 = L_c u$$

A rearrangement of terms gives us (5.1).

(4) $\Leftrightarrow$ (5):

From linear algebra, $R(L_c L_c^\star) = R(W_c)$, which establishes the equivalence of (4) and (5). ■

For linear time-invariant systems, we can provide an even more computationally efficient method of determining the controllability and observability of a system. This involves defining the **observability matrix** and *controllability matrix* of a system, as shown below.

Definition 5.12 (**Controllability Matrix, Observability Matrix**). *Let a linear time-invariant system $D = (A, B, C, D)$ be given. Define the* ***Controllability Matrix****, denoted by $\Sigma_C : \mathbb{R}^{n \times nn_i}$ and the* ***Observability Matrix****, denoted by $\Sigma_O : \mathbb{R}^{nn_o \times n}$, as follows:*

$$\Sigma_C = \begin{bmatrix} B & AB & \cdots & AB^{n-1} \end{bmatrix}$$

$$\Sigma_O = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix}$$

Remark. Notice that:

$$\begin{aligned} R(\Sigma_C) &\equiv R\left(\begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix} \right) \\ &= R(B) + R(AB) + \cdots + R(A^{n-1}B) \\ N(\Sigma_O) &\equiv N\left(\begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} \right) \\ &= N(C) \cap N(CA) \cap \cdots \cap N(CA^{n-1}) \end{aligned}$$

We are now ready to state a stronger version of the above theorem regarding controllability, as it pertains to linear time-invariant systems.

Theorem 5.13 (**Controllability of Linear Time-Invariant Systems**). *For a time-invariant system $R : \dot{x} = Ax + Bu$, the following statements are equivalent:*

1. *R is completely controllable (c.c.) on $[t_0, t_1]$.*
2. *For each $x_0 \in \mathbb{R}^n$, there exists some input $u_{[t_0,t_1]}$ that steers (x_0, t_0) to $(0, t_1)$.*
3. *For each $x_1 \in \mathbb{R}^n$, there exists some input $u_{[t_0,t_1]}$ that steers $(0, t_0)$ to (x_1, t_1).*
4. *The mapping $L_c : \mathcal{U}_{[t_0,t_1]} \to \mathbb{R}^n$ is surjective, i.e.:*
$$R(L_c) = \mathbb{R}^n$$
5. *The Controllability Grammian $W_c(t_0, t_1) \in \mathbb{R}^{n\times n}$ is invertible for each t, i.e.:*
$$\operatorname{rank}(W_c(t_0, t_1)) = n$$
for each t. In fact, $W_c(t_0, t_1) > 0$.
6. *$W_c(t_0, t_1) > 0$, and, if $\sigma(A) \in \mathbb{C}^-$, then $W_c(t_0, t_1)$ is the unique solution to the Lyapunov equation:*
$$AW_c + W_cA^\star = -BB^\star$$
7. *The Controllability matrix $\Sigma_C \in \mathbb{R}^{n\times nn_i}$ is of full row rank:*
$$\operatorname{rank}(\Sigma_C) = \operatorname{rank}\big(\begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix}\big) = n.$$
8. *For each $s \in \mathbb{C}$, the matrix $\begin{bmatrix} sI - A & B \end{bmatrix}$, known as the **Controllability Matrix Pencil** of (A, B), has full row rank, i.e.*
$$\operatorname{rank}\big(\begin{bmatrix} sI - A & B \end{bmatrix}\big) = n.$$
*This is known as the **Popov-Belovich-Hautus (PBH) test for controllability**. Since $sI - A$ lacks full row rank if and only if $s \in \sigma(A)$, the condition "for each $s \in \mathbb{C}$" can be replaced by the condition "for each $s \in \sigma(A)$" without any loss in generality.*
9. *For any polynomial $p(s)$ of degree n, there exists some $F \in \mathbb{R}^{n_i\times n}$ such that:*
$$\chi_{A+BF}(s) = p(s)$$

Proof.

(1) $\Leftrightarrow$ (2) $\Leftrightarrow$ (3) $\Leftrightarrow$ (4) $\Leftrightarrow$ (5):

See the proofs given for Theorem 5.11.

(5) $\Leftrightarrow$ (6)

If (5) holds, then $AW_c + W_cA^T = -BB^T$, with $\text{rank}(W_c(t_0, t_1)) = n$, so $W_c(t_0, t_1) > 0$. Moreover, by the lemma preceding the Time-Varying Lyapunov Lemma (Lemma 4.48), W_c is the unique solution to $AW_c + W_cA^T = -BB^T$.

Conversely, if $W_c(t_0, t_1) > 0$, then $\text{rank}(W_c(t_0, t_1)) = n$.

(5) $\Leftrightarrow$ (7)

For time invariant systems, $\Phi(t_1, \tau) = e^{(t_1-\tau)A}$, so the Controllability Grammian becomes:

$$W_c(t_0, t_1) = \int_{t_0}^{t_1} e^{(t_1-\tau)A} BB^\star e^{(t_1-\tau)A^\star}\, dt$$

Now, observe the equivalence of the following statements, which establishes the desired result by contradiction. Since $W_c(t_0, t_1) \geq 0$:

$$\begin{aligned}
&W_c(t_0, t_1) = \int_{t_0}^{t_1} e^{(t_1-\tau)A} BB^\star e^{(t_1-\tau)A^\star}\, d\tau > 0 \text{ is false,} \\
\iff &\exists x \neq 0 \ni x^\star W_c(t_0, t_1) x = \int_{t_0}^{t_1} |x^\star e^{(t_1-\tau)A} B|^2\, d\tau = 0, \\
\iff &\exists x \neq 0 \ni x^\star e^{\tau A} B = 0, \forall \tau \in [0, t_1 - t_0], \qquad (5.2) \\
\iff &\exists x \neq 0 \ni x^\star A^k B = 0, \forall k = 1, \cdots, n-1, \\
\iff &\exists x \neq 0 \ni x^\star \in LN(\underbrace{\begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix}}_{\equiv \Sigma_C}) \\
\iff &\text{rank}(\underbrace{\begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix}}_{\equiv \Sigma_C}) < n, \qquad (5.3)
\end{aligned}$$

Since $\Sigma_C \in \mathbb{R}^{n \times nn_i}$, we have $\Sigma_C \leq n$; this implies that exactly one of the statements "$\text{rank}(\Sigma_C) < n$" and "$\text{rank}(\Sigma_C) = n$" are true. The equivalence of the above statements thus forms a proof by contradiction for the claim "(5) $\Leftrightarrow$ (7)".

(7) $\Leftrightarrow$ (8)

(7) $\Rightarrow$ (8) can be straightforwardly demonstrated via proof by contradiction. If (8) fails to hold, then there exists some $\lambda \in \sigma(A)$, with corresponding left eigenvector $v^\star \neq 0$, such that $v^\star B = 0$. This implies that, for each $k = 1, \cdots, n-1$, we have:

$$v^\star A^k B = \lambda^k (v^\star B) = 0$$

and thus:

$$v^\star \Sigma_C = v^\star \begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix} = 0,$$

contradicting (7).

(8) $\Rightarrow$ (7) can also be shown via proof by contradiction, although not as straightforwardly. Suppose that (7) fails to hold, i.e.:

$$\text{rank}(\Sigma_C) = \text{rank}(\begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix}) < n$$

We wish to establish that (8) fails to hold, by considering a matrix representation of the Controllability Matrix Pencil $\begin{bmatrix} sI - A & B \end{bmatrix}$, that separates the controllable and uncontrollable subspaces of A. Notice that since (5) fails to hold, there exist collections of vectors:

$$\begin{aligned} \beta_{\Sigma_C} &\equiv \{v_1, \cdots, v_k\} \\ \beta' &\equiv \{v_{k+1}, \cdots, v_n\} \\ \beta &\equiv \beta_{\Sigma_C} \cup \beta' \\ &= \{v_1, \cdots, v_k, v_{k+1}, \cdots, v_n\} \end{aligned}$$

where $k < n$, and β_{Σ_C} and β form ordered bases for $R(\Sigma_C)$ and $\mathbb{R}^n$, respectively. To complete the picture, let $V \equiv \text{span}(\beta')$. We thus have:

$$\mathbb{R}^n = R(\Sigma_C) \oplus V_1$$

(Recall that $\oplus$ means direct sum.)

Now, notice that:

$$R(\Sigma_C) = R(B) + R(AB) + \cdots + R(A^{n-1}B)$$

is an A-invariant subspace containing $R(B)$. This allows us to simultaneously transform A into a block-upper-triangular form, while reducing several rows of B to 0. In particular, if we let T be the invertible square matrix whose columns are the vectors in the ordered basis $\mathcal{B}$ (placed in the same order as they appear in $\mathcal{B}$, then:

$$\begin{aligned} \tilde{A} &\equiv T^{-1}AT = \begin{bmatrix} \tilde{A}_{11} & \tilde{A}_{12} \\ O & \tilde{A}_{22} \end{bmatrix}, \\ \tilde{B} &\equiv T^{-1}B = \begin{bmatrix} \tilde{B}_1 \\ O \end{bmatrix}, \end{aligned}$$

In particular, the zero matrix block in $\tilde{A}$ arises from the invariance of $R(B) = \text{span}(\Sigma_C)$, whereas the zero matrix block in $\tilde{B}$ arises from the fact that, with respect to the ordered basis $\mathcal{B} = \mathcal{B}_{\Sigma_C} \cup \mathcal{B}$, only the first k coordinates of elements in $R(B)$ can be nonzero.

Now, notice that:

$$\begin{aligned} &\text{rank}(\begin{bmatrix} sI - A & B \end{bmatrix}) \\ =&\text{rank}\left(T \begin{bmatrix} sI - \tilde{A} & \tilde{B} \end{bmatrix} \begin{bmatrix} T^{-1} & O \\ O & I \end{bmatrix}\right) \\ =&\text{rank}(\begin{bmatrix} sI - \tilde{A} & \tilde{B} \end{bmatrix}) \\ =&\text{rank}\left(\begin{bmatrix} sI - \tilde{A}_{11} & -\tilde{A}_{12} & \tilde{B} \\ O & sI - \tilde{A}_{22} & O \end{bmatrix}\right) \end{aligned}$$

This shows that $\begin{bmatrix} sI - A & B \end{bmatrix}$ lacks full row rank (at least) whenever $s \in \sigma(\tilde{A}_{22})$, i.e. (7) does not hold.

(5) $\Leftrightarrow$ (9)

(See Lecture 20)

■

An analogous theorem exists for the equivalence of conditions that determine the observability of the system.

Theorem 5.14 (**Observability of Linear Time-Invariant Systems**). *For a time-invariant system $R : \dot{x} = Ax, y = Cx$, the following statements are equivalent:*

1. *R is completely observable (c.o.) on $[t_0, t_1]$.*
2. *The mapping $L_o : \mathcal{U}_{[t_0,t_1]} \to \mathbb{R}^n$ is injective, i.e.:*
$$N(L_c) = \{0\}$$
3. *The Observability Grammian $W_o(t_0, t_1) = (L_o^\star L_o)(t_0, t_1) \in \mathbb{R}^{n\times n}$ is invertible for each t, i.e.:*
$$\operatorname{rank}(W_o(t_0, t_1)) = n$$
for each t. In fact, $W_o(t_0, t_1) > 0$.
4. *$W_o(t_0, t_1) > 0$, and, if $\sigma(A) \in \mathbb{C}^-$, then $W_o(t_0, t_1)$ is the unique solution to the Lyapunov equation:*
$$A^\star W_o + W_o A = -C^\star C.$$
5. *The Observability matrix $\Sigma_O \in \mathbb{R}^{nn_o\times n}$ is of full column rank:*
$$\operatorname{rank}(\Sigma_O) = \operatorname{rank}\left(\begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix}\right) = n.$$
6. *For each $s \in \mathbb{C}$, the matrix $\begin{bmatrix} sI - A \\ C \end{bmatrix}$, known as the **Observability Matrix Pencil** of (A, B), has full column rank, i.e.*
$$\operatorname{rank}\left(\begin{bmatrix} sI - A \\ C \end{bmatrix}\right) = n.$$
*This is known as the **Popov-Belovich-Hautus (PBH) test for observability**. Since $sI - A$ lacks full column rank if and only if $s \in \sigma(A)$, the condition "for each $s \in \mathbb{C}$" can be replaced by the condition "for each $s \in \sigma(A)$" without any loss in generality.*

7. *For any polynomial $p(s)$ of degree n, there exists some $L \in \mathbb{R}^{n\times n_o}$ such that:*

$$\chi_{A+LC}(s) = p(s)$$

Proof. The proof of this theorem can be demonstrated in a manner analogous to the proof of Theorem 5.13. However, the results can also follow by observing the *duality* between controllabiity and observability, a concept captured by the following theorem. ■

Definition 5.15 (**Adjoint System** (Dual System))**.** *The **adjoint system** (or dual system) of the linear time-varying system:*

$$\Sigma : \begin{cases} \dot{x}(t) = A(t)x(t) + B(t)u(t) \\ y(t) = C(t)x(t) + D(t)u(t) \end{cases}$$

is defined as:

$$\overline{\Sigma} : \begin{cases} \dot{\overline{x}}(t) = -A^\star(t)\overline{x}(t) - C^\star(t)\overline{u}(t) \\ \overline{y}(t) = B^\star(t)\overline{x}(t) + D^\star(t)\overline{u}(t) \end{cases}$$

Note. Given a system S, the dual system of its dual system is S itself *with a change of sign for the state* ([4], Chapter 2, Comment 125, pg. 27):

$$\begin{aligned} -\dot{\tilde{x}}(t) &= A(t)\big(-\tilde{x}(t)\big) + B)(t)u(t), \\ \tilde{y}(t) &= C(t)\big(-\tilde{x}(t)\big) + D(t)u(t) \end{aligned}$$

Theorem 5.16 (**Properties of Adjoint Systems**)**.** *Let Σ and $\Sigma^\star$ be dual linear time-varying systems. Define $\Phi(t,t_0), L_c(t,t_0)$, and $L_o(t,t_0)$ to be the state transition matrix, Controllability Grammian, and Observability Grammian, of Σ, respectively, and define $\overline{\Phi}(t,t_0), \overline{L_c}(t,t_0)$, and $\overline{L_o}(t,t_0)$ similarly for $\overline{\Sigma}$. Then:*

1. $\overline{\Phi}(t,t_0) = \Phi^\star(t_0,t)$.
2. $\overline{L_c}(t_0,t) = \Phi^\star(t_0,t)\, L_o(t_0,t)\, \Phi(t_0,t)$.
3. $\overline{L_o}(t,t_0) = \Phi(t_0,t)\, L_c(t_0,t) \Phi^\star(t_0,t)$.

It follows that $\Sigma : (A,B,C,D)$ is observable if and only if $\overline{\Sigma} : (-A^\star, -C^\star, B^\star, D^\star)$ is controllable.

Proof.

1. Recall that $\Phi(t,t_0)$ and $\overline{\Phi}(t,t_0)$ are the unique time-dependent matrices satisfying:

$$\begin{aligned} \frac{d}{dt}\Phi(t,t_0) &= A(t)\,\Phi(t,t_0), \qquad & \Phi(t_0,t_0) &= I \\ \frac{d}{dt}\overline{\Phi}(t,t_0) &= -A^\star(t)\,\overline{\Phi}(t,t_0), \qquad & \overline{\Phi}(t_0,t_0) &= I \end{aligned}$$

Thus, to show that $\overline{\Phi}(t,t_0) = \Phi^\star(t_0,t)$, we must show that $\Phi^\star(t_0,t)$ satisfies the second differential equation and initial condition. We already know that $\Phi^\star(t_0,t) = I$; for the differential equation, we retrace the solution to Discussion 5, Problem 6, as follows:

$$\begin{aligned}
&\because\ I = \Phi(t,t_0)\,\Phi(t_0,t),\\
&\Rightarrow\ O = \frac{d}{dt}\Phi(t,t_0)\,\Phi(t_0,t) + \Phi(t,t_0)\,\frac{d}{dt}\Phi(t_0,t)\\
&\qquad = A(t)\,\Phi(t_0,t)\,\Phi(t_0,t) + \Phi(t,t_0)\,\frac{d}{dt}\Phi(t_0,t)\\
&\qquad = A(t) + \Phi(t,t_0)\,\frac{d}{dt}\Phi(t_0,t).
\end{aligned}$$

Rearranging terms, we have:

$$\begin{aligned}
\frac{d}{dt}\Phi(t_0,t) &= -\Phi(t_0,t)A(t),\\
\Rightarrow\ \frac{d}{dt}\Phi^\star(t_0,t) &= -A^\star(t)\Phi^\star(t_0,t).
\end{aligned}$$

where we have used the fact $\Phi(t,t_0)^{-1} = \Phi(t_0,t)$ in the last step.

2. We have:

$$\begin{aligned}
\overline{L_c}(t_0,t) &= \int_{t_0}^{t} \overline{\Phi}(t,\tau)\big(-C^\star(\tau)\big)\big(-C^\star(\tau)\big)^\star\,\overline{\Phi}^\star(t,\tau)\,d\tau\\
&= \int_{t_0}^{t} \Phi^\star(\tau,t)\,C^\star(\tau)\,C(\tau)\,\Phi(\tau,t)\,d\tau\\
&= \Phi^\star(t_0,t)\left(\int_{t_0}^{t} \Phi^\star(\tau,t_0)\,C^\star(\tau)\,C(\tau)\,\Phi(\tau,t_0)\,d\tau\right)\Phi(t_0,t)\\
&= \Phi^\star(t_0,t)\,L_o(t_0,t)\,\Phi(t_0,t)
\end{aligned}$$

3. We have:

$$\begin{aligned}
\overline{L_o}(t,t_0) &= \int_{t_0}^{t} \overline{\Phi}^\star(\tau,t_0)\,\big(B^\star(\tau)\big)^\star\,B^\star(\tau)\,\overline{\Phi}(\tau,t_0)\,d\tau\\
&= \int_{t_0}^{t} \Phi(t_0,\tau)\,B(\tau)B^\star(\tau)\,\Phi^\star(t_0,\tau)\,d\tau\\
&= \Phi(t_0,t)\left(\int_{t_0}^{t} \Phi(t,\tau)\,B(\tau)B^\star(\tau)\,\Phi^\star(t,\tau)\,d\tau\right)\Phi^\star(t_0,t)\\
&= \Phi(t_0,t)\,L_c(t_0,t)\Phi^\star(t_0,t)
\end{aligned}$$

Since $\Phi(t_0,t)\,\Phi(t,t_0) = I$, $\Phi(t_0,t)$ is invertible. Thus, $\overline{L_c}(t_0,t)$ is of full rank if and only if $L_o(t_0,t)$ is of full rank, and $\overline{L_o}(t_0,t)$ is of full rank if and only if $L_c(t_0,t)$ is of full rank. Part 5 of Theorem 5.13 therefore implies that $\Sigma : (A,B,C,D)$ is observable if and only if $\overline{\Sigma} : (-A^\star, -C^\star, B^\star, D^\star)$ is controllable..

■

This result allows us to intuitively (and, with a little work, rigorously) demonstrate the equivalence of Theorem 5.13 and Theorem 5.14. The corollary itself establishes Part 3 of Theorem 5.14, which states that a LTI system is observable if and only if its Obseravability Grammian has full rank. Part 6 of Theorem 5.14, which states the PBH test for observability also follows from the above corollary, by observing that the observability matrix pencil of an LTI system (A,B,C):

$$\begin{bmatrix} sI - A \\ C \end{bmatrix}$$

is of full column rank, for each $s \in \mathbb{C}$, if and only if the controllability matrix pencil of its adjoint system $(-A^\star, -C^\star, B^\star)$:

$$\begin{bmatrix} sI + A^\star & -C^\star \end{bmatrix},$$

for each $s \in \mathbb{C}$ (here, we take $D = D^\star = O$ in the definition of the adjoint system). Part 5 of Theorem 5.14, which concerns the column rank of the observability matrix, follows similarly from the above corollary.

Finally, in the event that the system is not completely controllable or completely observable, it is reasonable to ask whether or not a subset of the entire state space is controllable or observable. The following theorems address this concept (see Lecture 11 in Professor Tomlin's original notes).

Theorem 5.17. *For a time-invariant system $R : \dot{x} = Ax + Bu$, and each fixed t_0, t_1 with $t_0 < t_1$:*

$$R(W_c[t_0,t_1]) = R(\Sigma_C)$$

Proof. Since, for any linear map $A : V \to W$, we have $R(A) \oplus N(A^\star) = V$, we can instead verify that $N(W_c^\star[t_0,t_1]) = N(\Sigma_O{}^\star)$.

Observe that, for each $v \in \mathbb{R}^n$, the following statements are equivalent (i.e. for each $v \in \mathbb{R}^n$, they are either all true or all false):

$$\begin{aligned}
& v^\star W_c[t_1,t_0] = 0, \\
\iff & v^\star \left[\int_{t_0}^{t_1} e^{(t_1-\tau)A} B B^\star e^{(t_1-\tau)A^\star}\, d\tau \right] v = 0, \\
\iff & \int_{t_0}^{t_1} |v^\star e^{(t_1-\tau)A} B|^2\, d\tau = 0, \\
\iff & v^\star e^{(t_1-\tau)A} B = 0, \qquad \forall t \in [t_0,t_1], \\
\iff & v^\star \begin{bmatrix} B, AB, \cdots, A^{n-1}B \end{bmatrix} = 0.
\end{aligned}$$

The logic used in the forward direction of the above implications is straightforward; that of the backward direction involves the Cayley-Hamilton Theorem and the fact that $W_c[t_0, t_1]$ is positive semidefinite. ■

An analogous result holds for the observability of a system's output.

Theorem 5.18. *For each system $R : \dot{x} = Ax, y = Cx$, we have $N(W_0(t_0, t_1)) = N(\Sigma_O)$.*

The following theorem illustrates how $R(\Sigma_C)$ $(= R(W_c[t_0, t_1]))$ can be used to find states that are controllable in a system that is not completely controllable.

Theorem 5.19. *For a time-invariant system $R : \dot{x} = Ax + Bu$, if $x_1, x_2 \in \Sigma_C$, then there exists an input $u_{[t_0,t_1]}$ that steers $\dot{x} = Ax + Bu$ from (x_0, t_0) to (x_1, t_1).*

Proof. Since $x_0, x_1 \in R(W_c[t_0, t_1]) = R(L_c[t_0, t_1])$, and $R(W_c[t_0, t_1]) = R(\Sigma_O)$ is A-invariant, there exists some $u_{[t_0,t_1]}$, defined on $[t_0, t_1]$, such that:

$$\begin{aligned} L_c u(\cdot) &= x_1 - e^{(t-t_0)A}x_0, \\ \Rightarrow x_1 &= e^{(t-t_0)A}x_0 + L_c u(\cdot) \\ &= e^{(t-t_0)A}x_0 + \int_{t_0}^{t_1} e^{(t-\tau)A}Bu(\tau)\,d\tau, \end{aligned}$$

i.e. $u_{[t_0,t_1]}$ steers $\dot{x} = Ax + Bu$ from (x_0, t_0) to (x_1, t_1).

■

Finally, we discuss the controllability of a system over different time intervals.

Theorem 5.20. *Let $R : \dot{x} = Ax+Bu$ be a time-invariant system, and suppose $t_0 \le t_0' < t_1' \le t_1$. Then, if R is completely controllable on $[t_0', t_1']$, it is completely controllable on $[t_0, t_1]$.*

Proof. Suppose R is completely controllable on $[t_0', t_1']$. Let x_0, x_1 be arbitrarily given. We will show that a suitable input can be achieved by using zero input in the intervals $[t_0, t_0']$ and $[t_1', t_1]$, i.e.:

$$u_{[t_0,t_0']} = u_{[t_1',t_1]} = 0$$

Solving the differential equation $R : \dot{x} = Ax + Bu$ in the intervals $[t_0, t_0']$ and $[t_1', t_1]$, subject to the boundary conditions $x(t_0) = x_0, x(t_1) = x_1$, we have:

$$\begin{aligned} \dot{x} &= Ax + Bu = Ax \\ \Rightarrow x(t) &= \begin{cases} e^{(t-t_0)A}x_0, & t \in [t_0, t_0'], \\ e^{(t-t_1)A}x_1, & t \in [t_1', t_1] \end{cases} \end{aligned}$$

Now, since R is completely controllable on $[t_0', t_1']$, there exists an input $u_{[t_0',t_1']}$ that steers the system from $(e^{(t-t_0)A}x_0, t_0')$ to $(e^{(t_1'-t_1)A}x_1, t_1')$. In summary, we have thus demonstrated a sequence of three controls that steer the system as follows:

$$(x_0, t_0) \longrightarrow (e^{(t-t_0)A}x_0, t_0') \longrightarrow (e^{(t_1'-t_1)A}x_1, t_1') \longrightarrow (x_1, t_1)$$

■

Remark. The converse is not always true, i.e. a time-varying system that is controllable on a time interval may not necessarily be controllable on a strict subset of that time interval.

Consider, as a counterexample, the time-varying (but "piecewise time-invariant") system:

$$\dot{x} = A(t)x + B(t)u, \qquad x(0) = x_0.$$

where $A(t)$ and $B(t)$ are given by:

$$A(t) = 0, \qquad \forall t \geq 0,$$
$$B(t) = \begin{cases} 0, & t \in [0, 1), \\ 1, & t \geq 1. \end{cases}$$

It is straightforward to see that, when $t \geq 0$:

$$x(t) = \begin{cases} x_0, & t \in [0, 1), \\ x_0 + \int_1^t u(\tau)d\tau, & t \geq 1, \end{cases}$$

Then, for each $t \geq 1$, and arbitrarily fixed final state x_f, the constant control:

$$u \equiv \frac{x_f - x_0}{t - 1}$$

will drive the system from $(0, x_0)$ to (t, x_f). Thus, the system is controllable on $[0, t]$ for any $t > 1$.

However, the system is clearly not controllable on the smaller interval $[0, 1]$, since $x = x(0)$ regardless of our choice of input during that period of time.

5.3 Lectures 16, 17 Discussion

Example (**Discussion 11, Problem 2**). Show that, for any linear map $A : U \to V$, we have:

$$R(A^\star A) = R(A^\star)$$

Solution :

Below, we directly show that $R(A^\star A) \subset R(A^\star)$ and $R(A^\star) \subset R(A^\star A)$. Alternatively, one can also show that $N(AA^\star) = N(A^\star)$, and then invoke the fact that $V = R(A) \overset{\perp}{\oplus} N(A^\star)$ and $U = R(A^\star) \overset{\perp}{\oplus} N(A)$, and demonstrate that the orthogonal complement subspace of any fixed subspace is unique.

" $\subset$ " : Suppose $u \in R(A^\star A) \subset U$. Then there exists some $u' \in U$ such that $A^\star A u' = u$. This implies that u is the image of Au' through $A^\star$, so $u \in R(A^\star)$.

" $\supset$ " : Suppose $u \in R(A^\star)$. Then there exists some $v \in V$ such that $u = A^\star v$. Since $V = R(A) \overset{\perp}{\oplus} N(A^\star)$, there exists some $v_1 \in R(A)$ and $v_2 \in N(A^\star)$ such that $v = v_1 + v_2$. Moreover, since $v_1 \in R(A)$, there must exist some $u' \in U$ such that $v_1 = Au'$. Combining these facts, we have:

$$u = A^\star v = A^\star(v_1 + v_2) = A^\star A u' + A^\star v_2 = A^\star A u'.$$

Thus, $u \in R(A^\star A)$.

Example (**Discussion 11, Problem 3**). Consider the controllability and observability Grammians W_c, W_o of a linear time-invariant system (A, B, C) over the time period $[0, T]$. Prove that the $\sigma(W_c W_0)$ is invariant with respect to the similarity transformation:

$$(A, B, C) \qquad \to \qquad (TAT^{-1}, TB, CT^{-1}),$$

where T is non-singular.

Solution:

We have, for the Controllability Grammian of the transformed system:

$$\begin{aligned}
\overline{W_c}[0,t] &= \int_0^t e^{TAT^{-1}(t-\tau)} \cdot TB \cdot (TB)^\star \cdot e^{(TAT^{-1})^\star (t-\tau)}\, d\tau \\
&= T \left(\int_0^t e^{A(t-\tau)} T^{-1} \cdot TB \cdot B^\star T^\star \cdot (T^\star)^{-1} e^{A^\star (t-\tau)}\, d\tau \right) T^\star \\
&= T \left(\int_0^t e^{A(t-\tau)} BB^\star e^{A^\star (t-\tau)}\, d\tau \right) T^\star \\
&= T \cdot W_c[0,t] \cdot T^\star
\end{aligned}$$

Similarly, we can show that:

$$\overline{W_o}[0,t] = (T^\star)^{-1} \cdot W_c[0,t] \cdot T^{-1}$$

Thus, we have:

$$\overline{W_c W_o} = T(W_c W_o)^{-1} T^{-1},$$

so $\sigma(\overline{W_c W_o} = \sigma(W_c W_o)$.

Example (**Discussion 11, Problem 4**). A system (A, B) is one with dynamics of the form:

$$\dot{x} = Ax + Bu$$

For each of the following statements, provide either a proof or a counterexample.

1. Suppose the system (A, B) is controllable. Is the system (A^2, B) controllable?
2. Suppose the system (A^2, B) is controllable. Is the system (A, B) controllable?

Solution :

1. No. As a counterexample, let:

$$A^2 = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 & 1 \end{bmatrix}$$

Using the PBH test for controllability, we have:

$$\begin{bmatrix} sI - A & B \end{bmatrix} = \begin{bmatrix} s & -1 & 0 \\ 0 & s & 1 \end{bmatrix}$$
$$\begin{bmatrix} sI - A^2 & B \end{bmatrix} = \begin{bmatrix} s & 0 & 0 \\ 0 & s & 1 \end{bmatrix}$$

For each $s \in \mathbb{C}$, the controllability matrix pencil $\begin{bmatrix} sI - A & B \end{bmatrix}$ has full row rank; however, when $s = 0$, the matrix pencil $\begin{bmatrix} sI - A^2 & B \end{bmatrix}$ drops rank. We have thus established a counterexample such that (A, B) is controllable, but (A^2, B) is not.

2. The statement is true. To see this, observe that if (A^2, B) is controllable, then:

$$\text{rank}(\begin{bmatrix} B & A^2B & \cdots & A^{2(n-1)}B \end{bmatrix}) = n.$$

However, by the Cayley-Hamilton Theorem, for each $k \in \mathbb{N}$, the matrix A^k can be expressed as a linear combination of $\{I, A, \cdots, A^{n-1}B\}$, so:

$$R(\begin{bmatrix} B & A^2B & \cdots & A^{2(n-1)}B \end{bmatrix}) \subset R(\begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix}),$$
$$\Rightarrow n = \text{rank}(\begin{bmatrix} B & A^2B & \cdots & A^{2(n-1)}B \end{bmatrix}) \leq \text{rank}(\begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix})$$

But $\begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix} \in \mathbb{R}^{n \times nn_i}$, so its rank is at most n; the above inequality thus implies that its rank is in fact n. Thus, it has full row rank, and so (A, B) is controllable.

Example (**Discussion 11, Problem 5**). Let $L_1 = (A_1, b_1, c_1^T)$ and $L_2 = (A_2, b_2, c_2^T$ be completely controllable and completely observable single-input-single-output (SISO) systems. Discuss the controllability and observability of the systems:

$$L_3 = (A_3, B_3, C_3) = \left(\begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}, \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}, \begin{bmatrix} c_1^T & c_2^T \end{bmatrix} \right),$$
$$L_4 = (A_4, B_4, C_4) = \left(\begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}, \begin{bmatrix} b_1 & 0 \\ 0 & b_2 \end{bmatrix}, \begin{bmatrix} c_1^T & 0 \\ 0 & c_2^T \end{bmatrix} \right)$$

for the following two cases:

1. A_1, A_2 have no common eigenvalues.

2. A_1, A_2 have at least one common eigenvalues.

Solution :

We apply the PBH test for controllability and observability to both L_3 and L_4. Here, we assume $A_1 \in \mathbb{R}^{n_1 \times n_1}$ and $A_2 \in \mathbb{R}^{n_2 \times n_2}$:

$$L_3: \qquad \begin{bmatrix} sI - A & B_3 \end{bmatrix} = \begin{bmatrix} sI - A_1 & 0 & b_1 \\ 0 & sI - A_2 & b_2 \end{bmatrix}, \qquad \begin{bmatrix} sI \\ C_3 \end{bmatrix} = \begin{bmatrix} sI - A_1 & 0 \\ 0 & sI - A_2 \\ c_1^T & c_2^T \end{bmatrix},$$

$$L_4: \qquad \begin{bmatrix} sI - A & B_4 \end{bmatrix} = \begin{bmatrix} sI - A_1 & 0 & b_1 & 0 \\ 0 & sI - A_2 & 0 & b_2 \end{bmatrix}, \qquad \begin{bmatrix} sI \\ C_4 \end{bmatrix} = \begin{bmatrix} sI - A_1 & 0 \\ 0 & sI - A_2 \\ c_1^T & 0 \\ 0 & c_2^T \end{bmatrix},$$

Solution :

We discuss only the controllability of L_3 and L_4. Analogous results that hold for observability can be found through duality.

1. L_3:

   Observe the equivalence of the following statements:

   $$\begin{aligned} &L_3 \text{ is not controllable} \\ \Leftrightarrow &\exists\, s \in \mathbb{C}, \alpha^T = (\alpha_1, \cdots, \alpha_{n_1}), \beta^T = (\beta_1, \cdots, \beta_n), \text{ not both 0, such that:} \\ &\qquad 0 = \begin{bmatrix} \alpha^T & \beta^T \end{bmatrix} \begin{bmatrix} sI - A & B \end{bmatrix} = \begin{bmatrix} \alpha^T(sI - A_1) & \beta^T(sI - A_2) & \alpha^T b_1 + \beta^T b_2 \end{bmatrix} \end{aligned}$$

   If $\alpha = 0$ or $\beta = 0$, the other vector must be 0; in this case, the first inequality can be reduced to $\text{rank}(\begin{bmatrix} sI - A & b_1 \end{bmatrix}) < n$, or $\text{rank}(\begin{bmatrix} sI - A & b_2 \end{bmatrix}) < n$, contradicting the fact that L_1 and L_2 are both controllable. Thus, $\alpha \neq 0$ and $\beta \neq 0$.

   In summary, L_3 is uncontrollable if and only if there exist nonzero vectors $\alpha \in \mathbb{R}^{n_1}$ and $\beta \in \mathbb{R}^{n_2}$ such that:

   $$\begin{aligned} \alpha^T(sI - A_1) &= 0, \\ \beta^T(sI - A_2) &= 0, \\ \alpha^T b_1 + \beta^T b_2 &= 0. \end{aligned}$$

   This, in turn occurs if and only if $\sigma(A_1) \cap \sigma(A_2) \neq \phi$. Observe that, if this holds, we can always scale α^T and β^T to ensure the third equality holds.

   This result implies that, despite the complete controllability of L_1 and L_2, it is possible for L_3 to not be completely controllable, as long as A_1, A_2, b_1, b_2 satisfy certain criteria.

These criteria can be used to explicitly used to construct a counterexample; for instance, we could consider the rather trivial case:

$$A_1 = A_2 = 0, \qquad b_1 = b_2 = 1$$

where $n_1 = n_2 = 1$.

2. Repeating the above procedure, we arrive at the following four equalities instead:

$$\begin{aligned} \alpha^T(sI - A_1) &= 0, \\ \beta^T(sI - A_2) &= 0, \\ \alpha^T b_1 &= 0, \\ \beta^T b_2 &= 0, \end{aligned}$$

where α and β cannot both be 0. Again, the above logic can be used to argue that α and β, in fact, must both be nonzero. The first and third statements in the above list of equalities thus imply that L_1 is uncontrollable, while the second and fourth imply that L_2 is uncontrollable; both of these conclusions are contradictions to the assumptions given in the problem.

It follows that if L_1 and L_2 are completely controllable, the same must be true for L_4.

Example (**Discussion 11, Problem 6**). Consider the control system:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} -2 & 1 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(t)$$

Find some input $u : [0, 1] \to \mathbb{R}$ that takes the zero-state to $(1, 1)$ at time 1. (Hint: Try $u(t) = a_1 e^t + a_2 e^{2t}$.

Solution:

Taking the (single-sided) Laplace transform of the two differential equations, we have:

$$\begin{aligned} X_1(s) &= \frac{X_2(s)}{(s+2)} \frac{U(s)}{(s+1)(s+2)}, \\ X_2(s) &= \frac{U(s)}{(s+1)}. \end{aligned}$$

where $X_1(s), X_2(s), U(s)$ represent the Laplace transform of $x_1(t), x_2(t), u(t)$ respectively. Since $u(t)$ is a linear combination of e^t and e^{2t}, its Laplace transform must be of the form:

$$U(s) = \frac{a_1 s + a_0}{(s-1)(s-2)}$$

Substituting into the given differential equations, and applying Heaviside's method of partial fraction decomposition, we have:

$$\begin{aligned}
X_1(s) &= \frac{a_1 s + a_0}{(s+1)(s-1)(s-2)} \\
&= \frac{-\frac{1}{6}a_1 + \frac{1}{6}a_0}{s+1} + \frac{-\frac{1}{2}a_1 - \frac{1}{2}a_0}{s-1} + \frac{\frac{2}{3}a_1 + \frac{1}{3}a_0}{s-2} \\
X_2(s) &= \frac{a_1 s + a_0}{(s+1)(s+2)(s-1)(s-2)} \\
&= \frac{-\frac{1}{6}a_1 + \frac{1}{6}a_0}{s+1} + \frac{\frac{1}{6}a_1 - \frac{1}{12}a_0}{s+2} + \frac{-\frac{1}{6}a_1 - \frac{1}{6}a_0}{s-1} + \frac{\frac{1}{6}a_1 + \frac{1}{12}a_0}{s-2}
\end{aligned}$$

Taking the inverse Laplace transform, we have:

$$\begin{aligned}
x_1(t) &= \left(-\frac{1}{6}a_1 + \frac{1}{6}a_0\right)e^{-t} + \left(-\frac{1}{2}a_1 - \frac{1}{2}a_0\right)e^{t} + \left(\frac{2}{3}a_1 + \frac{1}{3}a_0\right)e^{2t} \\
x_2(t) &= \left(-\frac{1}{6}a_1 + \frac{1}{6}a_0\right)e^{-t} + \left(\frac{1}{6}a_1 - \frac{1}{12}a_0\right)e^{-2t} + \left(\frac{1}{6}a_1 - \frac{1}{6}a_0\right)e^{t} + \left(\frac{1}{6}a_1 + \frac{1}{12}a_0\right)e^{2t}
\end{aligned}$$

Substituting $t = 1$ and collecting terms, we have:

$$\begin{aligned}
1 &= \left(-\frac{1}{6}e^{-1} - \frac{1}{2}e + \frac{2}{3}e^{2}\right)a_1 + \left(\frac{1}{6}e^{-1} - \frac{1}{2}e + \frac{1}{3}e^{2}\right)a_0 \\
&\approx 3.506a_1 + 1.165a_0, \\
1 &= \left(-\frac{1}{6}e^{-1} + \frac{1}{6}e^{-2} + \frac{1}{6}e + \frac{1}{6}e^{2}\right)a_1 + \left(\frac{1}{6}e^{-1} - \frac{1}{12}e^{-2} - \frac{1}{6}e + \frac{1}{12}e^{2}\right)a_0 \\
&\approx 1.646a_1 + 0.213a_0.
\end{aligned}$$

Rewriting the above equations in a matrix form, we find that:

$$\begin{bmatrix} a_1 \\ a_0 \end{bmatrix} \approx \begin{bmatrix} 3.506 & 1.165 \\ 1.646 & 0.213 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -0.182 & 0.995 \\ 1.406 & -2.995 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0.813 \\ -1.589 \end{bmatrix}$$

Substituting back to $U(s)$, we have:

$$\begin{aligned}
U(s) &= \frac{a_1 s + a_0}{(s-1)(s-2)} = \frac{-a_1 - a_0}{s-1} + \frac{2a_1 + a_0}{s-2} = \frac{0.776}{s-1} + \frac{0.038}{s-2}, \\
\Rightarrow u(t) &= 0.776e^{t} + 0.038e^{2t}
\end{aligned}$$

Remark.

1. The initial conditions $(x_1, x_2)(0) = (0, 0)$ and final conditions $(x_1, x_2)(1) = (1, 1)$ technically imply that we need to solve a system of four linear equations. However, by taking the (single-sided) Laplace transform, we automatically take into account the initial conditions, leaving us with the final conditions. This is the reason we only required two variables, a_1, and a_2.

2. In fact, other inputs, such as those of the form $u(t) = a_1 t + a_0$ or $u(t) = a_1 e^t + a_2 t$ would also suffice.

5.4 Lecture 18

In this section, we will discuss *observers* and *least L^2-norm inputs*.

Observers:

Consider the following linear time-variant system without inputs:

$$\begin{aligned} \dot{x} &= A(t)x, \\ y &= C(t)x, \end{aligned}$$

In the following discussion, we assume that $W_0[t_0, t_1]$ is completely observable, i.e. $W_0^{-1}[t_0, t_1]$ is well-defined. In the case that this is not true, we can simply replace $W_0^{-1}[t_0, t_1]$ with $W_0^{\dagger}[t_0, t_1]$, the *pseudo-inverse* of $W_0[t_0, t_1]$ (see, for instance [5], Section 6.7, pgs. 413-417 for details).

By the definition of the observability map:

$$y(\cdot) = L_0 x_0 = C(\cdot)\Phi(\cdot, t_0)x_0.$$

The minimum least-square estimate of x_0 from y_0 is thus:

$$\begin{aligned} x_0 &= \left(L_0^{\star}[\cdot, t_0] \cdot L_0[\cdot, t_0]\right)^{-1} L_0^{\star}[\cdot, t_0])y(\cdot) \\ &= W_0^{-1}[t_0, t] \cdot \int_{t_0}^{t} \Phi^{\star}(\tau, t_0)C^{\star}(\tau)y(\tau)\, d\tau \end{aligned}$$

Define $\hat{x}(t)$ to be the *optimal estimate of $x(t)$ based on $\{y(\tau) : t_0 \leq \tau \leq t\}$*. Then:

$$\begin{aligned} \hat{x}(t_0) &= W_0^{-1}[t, t_0]L_0^{\star}[t, t_0]y(\cdot) \\ \Rightarrow\ \hat{x}(t) &= \Phi(t, t_0)W_0^{-1}[t, t_0]L_0^{\star}[t, t_0]y(\cdot) \end{aligned}$$

Below, we wish to establish a recursive relation for $\hat{x}(t)$, to characterize the difference between the time evolution of the system state $x(t)$ and that of the optimal estimate $\hat{x}(t)$. Applying the

product rule for differentiation, we have:

$$\begin{aligned}
\frac{d}{dt}\hat{x}(t) &= A(t)\,\Phi(t,t_0)\,W_0^{-1}[t,t_0]\cdot L_0^\star[t,t_0]y(\cdot) \\
&\qquad + \Phi(t,t_0)\left(\frac{d}{dt}W_0^{-1}\right)[t,t_0]\cdot L_0^\star[t,t_0]y(\cdot) \\
&\qquad + \Phi(t,t_0)\,W_0^{-1}[t,t_0]\,\Phi^\star(t,t_0)\,C^\star(t)\,y(t) \\
&= A(t)\hat{x}(t) \\
&\qquad + \Phi(t,t_0)\left(-W_0^{-1}\dot{W}_0W_0^{-1}\right)[t,t_0]\cdot L_0^\star[t,t_0]y(\cdot) \\
&\qquad + \Phi(t,t_0)\,W_0^{-1}[t,t_0]\,\Phi^\star(t,t_0)\,C^\star(t)\,y(t) \\
&= A(t)\hat{x}(t) \\
&\qquad - \Phi(t,t_0)(W_0^{-1}[t,t_0]\cdot\Phi^\star(t,t_0)C^\star(t)C(t)\underbrace{\Phi(t,t_0)\cdot W_0^{-1}[t,t_0])\cdot L_0^\star[t,t_0]y(\cdot)}_{=\hat{x}(t)} \\
&\qquad + \Phi(t,t_0)\,W_0^{-1}[t,t_0]\,\Phi^\star(t,t_0)\,C^\star(t)\,y(t) \\
&= A(t)\hat{x}(t) + \underbrace{\Phi(t,t_0)W_0^{-1}[t,t_0]\Phi^\star(t,t_0)}_{P(t)\text{: Kalman Filter gain}}C^\star(t)\,\underbrace{\big(y(t)-C(t)\hat{x}(t)\big)}_{\text{observed error}}
\end{aligned}$$

In other words, the optimal state estimate evolves with a rate of change differing from that of the original system by an amount proportional to the state estimation error. This proportionality is described by the *Kalman Filter gain* $P(t)$, a positive semi-definite matrix.

The evolution of $P(t)$ can be characterized as follows:

$$\begin{aligned}
\dot{P}(t) &= A(t)\,P(t) + P(t)\,A^\star(t) \\
&\qquad - \Phi(t,t_0)\,W_0^{-1}[t,t_0]\,\Phi^\star(t,t_0)\,C^\star(t)\,C(t)\,\Phi(t,t_0)\,W_0^{-1}[t,t_0]\,\Phi^\star(t,t_0) \\
&= A(t)\,P(t) + P(t)\,A^\star(t) - P(t)C^\star(t)C^\star(t)C(t)P(t)
\end{aligned}$$

Notice that since $W_0[t,t_0]=0$, the Kalman filter gain is, in fact, not well-defined at $t=0$. Thus, instead of directly observing the evolution of $P(t)$, we instead observe the evolution of $Q(t)\equiv P^{-1}(t)$:

$$\dot{Q}(t) = -Q(t)\,A(t) - A^\star(t)\,Q(t) + C^\star(t)C(t)$$

where $Q(0)=0$, since $W_0[t,t_0]=0$.

If there are inputs to the linear system, we simply subtract the effect of the input on the system from the total observation:

$$\begin{aligned}
y(t) &= C(t)\,\Phi(t,t_0)x_0 + \int_{t_0}^{t} C(t)\,\Phi(t,\tau)\,B(\tau)\,u(\tau)\,d\tau + D(t)\,u(t) \\
\Rightarrow\ z(t) &\equiv y(t) - \int_{t_0}^{t} C(t)\,\Phi(t,\tau)\,B(\tau)\,u(\tau)\,d\tau - D(t)\,u(t) \\
&= C(t)\,\Phi(t,t_0)x_0
\end{aligned}$$

The above analysis follows; we simply replace each (total observation) $y(t)$ with (observation characterizing the evolution of $x(t)$) $z(t)$.

Least L_2 norm input to steer from x_0 to x_1:

Consider the linear time-variant dynamical system:

$$\begin{aligned} \dot{x}(t) &= A(t)x + B(t)u, \\ y(t) &= C(t)x + D(t)u. \end{aligned}$$

We wish to solve for the optimal control, in the least L_2-norm sense, that drives the system from (x_0, t_0) to (x_1, t_1):

$$\begin{aligned} x(t_1) &= \Phi(t_1, t_0)\, x(t_0) + \int_{t_0}^{t_1} \Phi(t_1, \tau)\, B(\tau)\, u(\tau)\, d\tau \\ &= \Phi(t_1, t_0)\, x(t_0) + L_c[t, t_0]\, u(\cdot) \\ \Rightarrow\ L_c[t, t_0]\, u(\cdot) &= x_1 - \Phi(t_1, t_0) x_0 \end{aligned}$$

In the event that L_c is not invertible, the (unique) optimal solution for $u(\cdot)$ must lie in $N(L_c)^\perp = R(L_c^\star)$, i.e. there must exist some $w \in \mathbb{R}^n$ such that $L_c^\star[t, t_0]\, w = u(\cdot)$. We thus have:

$$\begin{aligned} (L_c\, L_c^\star)[t, t_0]\, w_0 &= x_1 - \Phi(t_1, t_0) x_0 \\ \Rightarrow w_0 &= \big(L_c\, L_c^\star\big)^{-1}[t, t_0]\big(x_1 - \Phi(t_1, t_0) x_0\big) \\ \Rightarrow u(\cdot) &= L_c^\star[t, t_0]\big(L_c\, L_c^\star\big)^{-1}[t, t_0]\big(x_1 - \Phi(t_1, t_0) x_0\big) \end{aligned}$$

5.5 Lecture 19

In Lecture 17, we established that a linear time-invariant system (A, B, C, D), with n-dimensional state space Σ, is completely controllable if and only if:

$$R(\Sigma_C) \equiv R(\begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix}) = \mathbb{R}^n$$

and completely observable if and only if:

$$N(\Sigma_O) \equiv N\left(\begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix}\right) = \{0\}$$

We now wish to quantitatively investigate whether a linear time-invariant system (A, B, C, D) can be partially controllable or observable if it is not completely so.

In the most general case, it is possible that (A, B, C, D) is neither completely controllable nor completely observable. In this case, we wish to categorize states as either "controllable," "observable," both, or neither. To that end, consider the definitions below.

Definition 5.21 (Reachable from 0, Reachable, Unobservable, Indistinguishable States).

1. *Given an initial time and state (x_0, t_0), the state x_1 is **reachable from 0 on** $[t_0, t_1]$ are those for which there exists some input $u(\cdot)_{[t_0,t_1]}$ such that:*

$$x_1 = L_c[t_0, t_1]\, u(\cdot),$$

i.e. $x_1 \in R(\Sigma_C)$.

2. *Given an initial time and state (x_0, t_0), the state x_1 is **reachable on** $[t_0, t_1]$ are those for which there exists some initial state x_0 and input $u(\cdot)_{[t_0,t_1]}$ such that:*

$$x_1 = L_c[t_0, t_1]\, u(\cdot) + e^{(t-t_0)A} x_0,$$

i.e. $x_1 \in R(\Sigma_C) + e^{(t-t_0)A} x_0$.

3. *Given initial and final times t_0, t_1, respectively, the state x_0 is said to be **unobservable at** t_1 **from** (x_0, t_0) if the corresponding zero input response is 0, i.e.:*

$$L_0[t_0, t_1]\, x_0 = 0$$

i.e. $x_0 \in N(\Sigma_O)$.

4. *Given initial and final times t_0, t_1, respectively, the states x_{01} and x_{02} are said to be **indistinguishable at** t_1 **from** (x_0, t_0) if the corresponding zero input responses are the same, i.e.:*

$$L_0[t_0, t_1]\, x_{01} = L_0[t_0, t_1]\, x_{02},$$

i.e. $x_{02} \in N(\Sigma_O) + x_{01}$.

Next, we wish to establish certain properties of controllable and observable subspaces—namely, the fact that they are A-invariant.

Theorem 5.22. *Given a linear time-invariant dynamical system (A, B, C, D):*

1. *$R(\Sigma_C)$ is the smallest A-invariant subspace of Σ that contains $R(B)$.*
2. *$N(\Sigma_O)$ is the largest A-invariant subspace of Σ that contains $N(C)$.*

Proof.

1. There are four claims to verify:
   - $R(\Sigma_C)$ is a subspace of Σ.
   - $R(\Sigma_C)$ contains $R(B)$.
   - $R(\Sigma_C)$ is A-invariant.
   - $R(\Sigma_C)$ is the smallest subspace of Σ that satisfies the above two properties, i.e. any other subspace satisfying the above properties contains $R(\Sigma_C)$.

   The first two claim are true by the definition $R(\Sigma_C = R(B)+R(AB)+\cdots+R(A^{n-1}B)$.

   To verify the third claim, let us adopt the following notation, for any mapping $A : V \to W$ and subset $S \subset V$:

$$A(S) \equiv \{Ax | x \in S\} = R(A|_S)$$

   The third claim essentially states that $A(R(\Sigma_C)) \subset R(\Sigma_C)$. This follows from the fact that:

$$\begin{aligned} A(R(B)) &\equiv R(A\Sigma_C) \\ &= \begin{bmatrix} R(AB) & R(A^2B) & \cdots & R(A^nB) \end{bmatrix} \\ &\subset \begin{bmatrix} R(B) & R(AB) & \cdots & R(A^{n-1}B) \end{bmatrix} \qquad (5.4) \\ &\subset \begin{bmatrix} R(B) & R(AB) & R(A^2B) & \cdots & R(A^{n-1}B) \end{bmatrix} \qquad (5.5) \\ &= R(B), \end{aligned}$$

   where (5.5) follows from Cayley-Hamilton Theorem, which implies that A^n is a linear combination of $I, B, \cdots, A^{n-1}B$. This in turn implies that:

$$R(A^nB) \subset \begin{bmatrix} R(B) & R(AB) & R(A^2B) & \cdots & R(A^{n-1}B) \end{bmatrix} = R(B)$$

   It remains to demonstrate the fourth statement. Let V be any A-invariant subspace of Σ that contains $R(B)$. Then, since V is A-invariant, it must also contain:

$$A^k(R(B)) = R(A^k(B))$$

   for any $k \in \mathbb{N}$. It thus contains:

$$R(\Sigma_C) = R\Big(\begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix} \Big),$$

   completing the proof.

2. Again, we have four statements to verify:

   - $N(\Sigma_O)$ is a subspace of Σ.
   - $N(\Sigma_O)$ contains $N(C)$.
   - $N(\Sigma_O)$ is A-invariant.
   - $N(\Sigma_O)$ is the largest subspace of Σ that satisfies the above two properties, i.e. any other subspace satisfying the above properties must be contained in $N(\Sigma_O)$.

   Again, the first two claims follows from the definition of $N(\Sigma_O)$:

$$N(\Sigma_O) = N(C) \cap N(CA) \cap \cdots \cap N(CA^{n-1})$$

   To show that the third claim holds, suppose $x \in N(\Sigma_O)$. Then:

$$\begin{aligned} &Cx = CAx = \cdots = CA^{n-1}x = 0, \\ \Rightarrow &C(Ax) = C(A^2x) = \cdots = CA^n x = 0, \end{aligned}$$

   i.e. $Ax \in N(\Sigma_O)$, since, by the Cayley-Hamilton Theorem, A^n is a linear combination of $I, A, \cdots, A^{n-1}$. This shows that $A(N(\Sigma_O)) \subset N(\Sigma_O)$, i.e. $N(\Sigma_O)$ is A-invariant.

   To verify the fourth claim, let W be any A-invariant subspace of Σ that contains $N(C)$. Thus, $C(W) = \{0\}$, and since W is A-invariant, i.e. $A^k(W) \subset W$ for each $k \in \mathbb{N}$, we have:

$$C(A^k W) \subset C(W) = \{0\},$$

   i.e. $W \subset N(CA^k)$ for each $k \in \mathbb{N}$. Thus:

$$W \subset N(A) \cap N(CA) \cap \cdots \cap N(CA^{n-1} = N(\Sigma_O),$$

   completing the proof.

■

The above proof allows us to decompose the state space Σ according to whether the states are reachable from 0, controllable, both, or neither. Now, let us define:

$$\begin{aligned} \Sigma_C &\equiv R(\Sigma_C) \\ \Sigma_{O'} &\equiv N(\Sigma_O) \end{aligned}$$

we have $\Sigma_C \neq \mathbb{R}^n$ and $\Sigma_{O'} \neq \{0\}$, so there exist non-zero subspaces of the state space Σ, denoted as $\Sigma_{C'}, \Sigma_O$ (not unique), such that:

$$\Sigma = \Sigma_C + \Sigma_{C'} = \Sigma_{O'} + \Sigma_O,$$

Intuitively speaking, $\Sigma_C, \Sigma_{C'}, \Sigma_{O'}, \Sigma_O$ correspond to the controllable, uncontrollable, observable, and unobservable subspaces of the state space, respectively.

Although the state subspaces $\Sigma_C \equiv R(O)$ and $\Sigma_{O'} \equiv N(\Sigma_O)$ are uniquely given by the system parameters A, B, C, D, the state subspaces $\Sigma_{C'}$ and Σ_O are not. This is because, while it is indeed true that any state in $\Sigma_C, \Sigma_{C'}, \Sigma_{O'}, \Sigma_O$ would be controllable from 0, uncontrollable from 0, observable, and unobservable on $[t_0, t_1]$, respectively, it is not true that states uncontrollable from 0 are restricted to $\Sigma_{C'}$, nor is it true that the observable states are restricted to Σ_O. This, in turn, is because the sum of a controllable-from-0 and an uncontrollable-from-0 state is a state that is not controllable from 0; similarly, the sum of an unobservable and an observable state gives an observable state (as can be verified by substituting into the above definitions).

Now, we divide the state space Σ into subspaces based on controllability and observability:

$$\begin{aligned}\Sigma_{CO} &= \Sigma_C \cap \Sigma_O,\\ \Sigma_{CO'} &= \Sigma_C \cap \Sigma_{O'},\\ \Sigma_{C'O} &= \Sigma_{C'} \cap \Sigma_O,\\ \Sigma_{C'O'} &= \Sigma_{C'} \cap \Sigma_{O'},\end{aligned}$$

where the subscripts C, O, C', O' indicate that the subspace is controllable, observable, uncontrollable, or unobservable, respectively. Among these four subspaces, however, only $\Sigma_{CO'} = \Sigma_C \cap \Sigma_{O'}$ is uniquely defined. However, regardless of our choice of $\Sigma_{CO}, \Sigma_{C'O}$, and $\Sigma_{C'O'}$, we always have:

$$\Sigma = \Sigma_{CO} \oplus \Sigma_{CO'} \oplus \Sigma_{C'O} \oplus \Sigma_{C'O'}$$

Below, we find suitable matrix representations for the parameters of the systems dynamics, A, B, C, D, to illuminate the controllability and observability of the system.

Kalman Decomposition Theorem

Let $\mathcal{B}_{CO}, \mathcal{B}_{CO'}, \mathcal{B}_{C'O}, \mathcal{B}_{C'O'}$ be ordered bases for $\Sigma_{CO}, \Sigma_{CO'}, \Sigma_{C'O}, \Sigma_{C'O'}$, respectively. Then:

$$\begin{aligned}\mathcal{B}_C &\equiv \mathcal{B}_{CO} \cup \mathcal{B}_{CO'},\\ \mathcal{B}_{O'} &\equiv \mathcal{B}_{CO'} \cup \mathcal{B}_{C'O'},\\ \mathcal{B} &\equiv \mathcal{B}_{CO} \cup \mathcal{B}_{CO'} \cup \mathcal{B}_{C'O} \cup \mathcal{B}_{C'O'},\end{aligned}$$

with the union taken in that order, are ordered bases for Σ_C, $\Sigma_{O'}$, and Σ, respectively.

Let $[A]_{\mathcal{B}}$, $[B]_{\mathcal{B}}$, and $[C]_{\mathcal{B}}$ be the matrix representations of A, B, C, respectively, with respect to the ordered basis $\mathcal{B}$. Since $\Sigma_C \equiv R(\Sigma_C)$ is A-invariant, the 2nd Representation Theorem (Theorem 4.10) implies that these matrix representations are of the form:

$$[A]_{\mathcal{B}} = \begin{bmatrix} A_{11} & A_{12} & A_{13} & A_{14} \\ A_{21} & A_{22} & A_{23} & A_{24} \\ O & O & A_{33} & A_{34} \\ O & O & A_{43} & A_{44} \end{bmatrix}, \qquad [B]_{\mathcal{B}} = \begin{bmatrix} B_1 \\ B_2 \\ O \\ O \end{bmatrix},$$
$$[C]_{\mathcal{B}} = \begin{bmatrix} C_1 & C_2 & C_3 & C_4 \end{bmatrix},$$

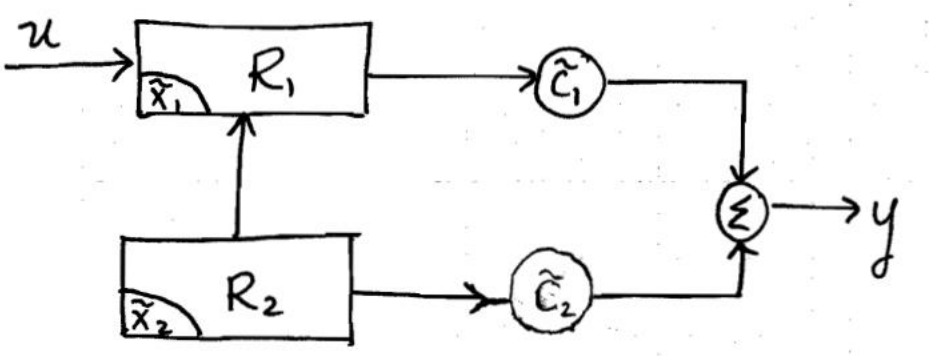
Similarly, since $\Sigma_{O'} \equiv N(\Sigma_O)$ is A-invariant, these matrix representations must also be of the form:

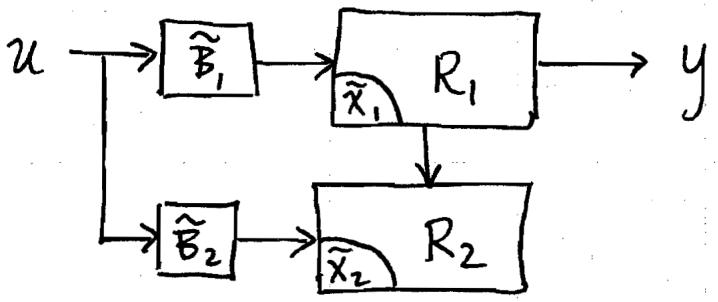
$$[A]_{\mathcal{B}} = \begin{bmatrix} A_{11} & O & A_{13} & O \\ A_{21} & A_{22} & A_{23} & A_{24} \\ A_{31} & O & A_{33} & O \\ A_{41} & A_{42} & A_{43} & A_{44} \end{bmatrix}, \qquad [B]_{\mathcal{B}} = \begin{bmatrix} B_1 \\ B_2 \\ B_3 \\ B_4 \end{bmatrix},$$
$$[C]_{\mathcal{B}} = \begin{bmatrix} C_1 & O & C_3 & O \end{bmatrix},$$

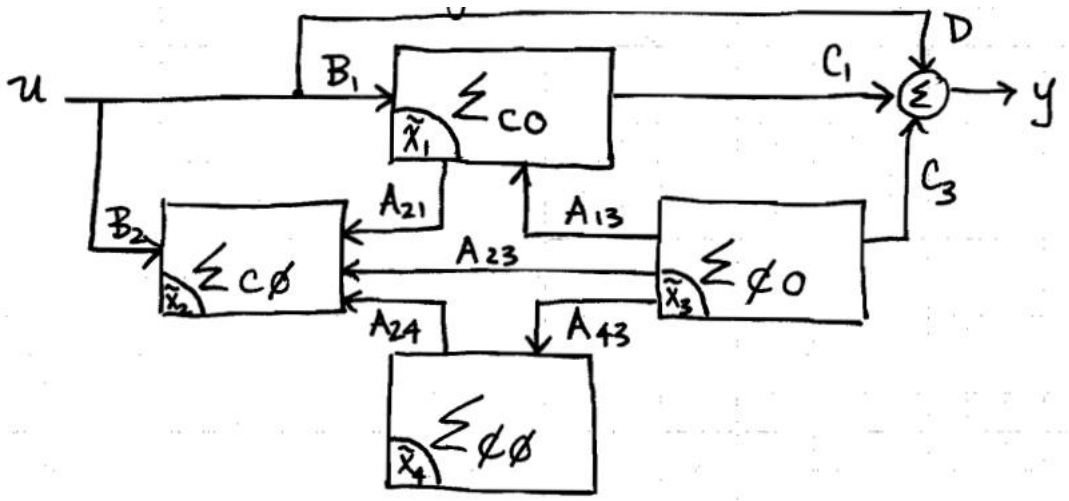
Combining the above two facts, we have:

$$[A]_{\mathcal{B}} = \begin{bmatrix} A_{11} & O & A_{13} & O \\ A_{21} & A_{22} & A_{23} & A_{24} \\ O & O & A_{33} & O \\ O & O & A_{43} & A_{44} \end{bmatrix}, \qquad [B]_{\mathcal{B}} = \begin{bmatrix} B_1 \\ B_2 \\ O \\ O \end{bmatrix},$$
$$[C]_{\mathcal{B}} = \begin{bmatrix} C_1 & O & C_3 & O \end{bmatrix},$$

The controllability and observability of these four subspaces can be expressed using the following diagrams:







Finally, we will show that the transfer function of a system only depends on the subspace of the state space that is both observable and controllable. However, to do so, we must first establish the following formula regarding the inverse of block-upper-triangular matrices. We invoke a simplified case of Schur Decomposition.

Lemma 5.23 (**Schur Complement, Simplified Form**). *Consider* $A, B \in \mathbb{R}^{(n+k)\times(n+k)}$*:*

$$A = \begin{bmatrix} A_{11} & A_{12} \\ O & A_{22} \end{bmatrix}, \qquad B = \begin{bmatrix} B_{11} & O \\ B_{21} & B_{22} \end{bmatrix},$$

where $A_{11}, B_{11} \in \mathbb{R}^{n\times n}$*,* $A_{12} \in \mathbb{R}^{n\times k}$*,* $B_{21} \in \mathbb{R}^{k\times n}$*, and* $A_{22}, B_{22} \in \mathbb{R}^{k\times k}$*. If* A_{11} *and* A_{12} *are invertible, then so is* A*, and:*

$$A^{-1} = \begin{bmatrix} A_{11}^{-1} & -A_{11}^{-1}A_{12}A_{22}^{-1} \\ 0 & A_{22}^{-1} \end{bmatrix}, \qquad B^{-1} = \begin{bmatrix} B_{11}^{-1} & O \\ -B_{22}^{-1}B_{21}B_{11}^{-1} & B_{22}^{-1} \end{bmatrix} \tag{5.6}$$

Proof. We can straightforwardly verify that (5.6) gives the correct expression for the left inverse of A, as shown below:

$$\begin{bmatrix} A_{11}^{-1} & -A_{11}^{-1}A_{12}A_{22}^{-1} \\ 0 & A_{22}^{-1} \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ O & A_{22} \end{bmatrix} = \begin{bmatrix} I_n & O \\ O & I_k \end{bmatrix} = I_{n+k}.$$

Since A is a square matrix, the existence of its left inverse implies that its right inverse must also exist, and equal its left inverse.

For B, we simply take the transpose of A, the apply the equation derived for A^{-1}.

■

We are now ready to show that the transfer function of a system only depends on the subspace of the state space that is both observable and controllable.

Theorem 5.24. *Consider an LTI system (A, B, C) with dynamics:*

$$\begin{aligned}\dot{x} &= Ax + Bu,\\ y &= Cx\end{aligned}$$

and Kalman decomposition:

$$[A]_{\mathcal{B}} = \begin{bmatrix} A_{11} & O & A_{13} & O \\ A_{21} & A_{22} & A_{23} & A_{24} \\ O & O & A_{33} & O \\ O & O & A_{43} & A_{44} \end{bmatrix}, \qquad [B]_{\mathcal{B}} = \begin{bmatrix} B_1 \\ B_2 \\ O \\ O \end{bmatrix},$$
$$[C]_{\mathcal{B}} = \begin{bmatrix} C_1 & O & C_3 & O \end{bmatrix},$$

Then the transfer function of (A, B, C) is:

$$H(s) \equiv C(sI - A)^{-1}B = C_1(sI - A_{11})^{-1}B_1$$

Proof. First, we observe that change of coordinates does not affect the controllability, observability, or transfer function of a system. Thus, it suffices to verify that

$$[C]_{\mathcal{B}}(sI - [A]_{\mathcal{B}})^{-1}[B]_{\mathcal{B}} = C_1(sI - A_{11})^{-1}B_1$$

Define $K(s) = (sI - [A]_{\mathcal{B}})^{-1}$. Since Schur Decomposition implies that the inverse of an invertible block-upper-triangular (respectively, block-lower-triangular) matrix must also be block-upper-triangular (respectively, block-lower-triangular), K must have a form similar to A, i.e.:

$$K = \begin{bmatrix} K_{11} & O & K_{13} & K_{14} \\ K_{21} & K_{22} & K_{23} & K_{24} \\ O & O & K_{33} & O \\ O & O & K_{43} & K_{44} \end{bmatrix}$$

Thus, we have:

$$\begin{aligned} H(s) &= [C]_{\mathcal{B}}(sI - [A]_{\mathcal{B}})^{-1}[B]_{\mathcal{B}} = [C]_{\mathcal{B}}K(s)[B]_{\mathcal{B}} \\ &= \begin{bmatrix} C_1 & O & C_3 & O \end{bmatrix} \begin{bmatrix} K_{11} & O & K_{13} & K_{14} \\ K_{21} & K_{22} & K_{23} & K_{24} \\ O & O & K_{33} & O \\ O & O & K_{43} & K_{44} \end{bmatrix} (s) \begin{bmatrix} B_1 \\ B_2 \\ O \\ O \end{bmatrix} \\ &= C_1(sI - A_{11})^{-1}B_1 \end{aligned}$$

where we have again used the Schur Complement to show that $K_{11}(s) = (sI - A_{11})^{-1}$.

■

We conclude our discussion by pointing out several important properties of the Kalman canonical form.

Theorem 5.25.

1. *Suppose a system $R : \dot{x} = Ax + Bu$ has controllability matrix Σ_C with rank $n_c < n$, and Kalman decomposition of the form:*

$$[A]_{\mathcal{B}} = \begin{bmatrix} A_{11} & A_{12} \\ O & A_{22} \end{bmatrix}, \qquad [B]_{\mathcal{B}} = \begin{bmatrix} B_1 \\ O \end{bmatrix}$$

where $\mathcal{B}$ is an ordered basis for $\mathbb{R}^n$, the first n_c columns of which form a basis for Σ_C. Then (A_{11}, B_1) is completely controllable.

2. *Suppose a system $R : \dot{x} = Ax$, $y = Cx$ has observability matrix Σ_O with rank $n_o < n$, and Kalman decomposition of the form:*

$$[A]_{\mathcal{B}} = \begin{bmatrix} A_{11} & O \\ A_{21} & A_{22} \end{bmatrix}, \qquad [C]_{\mathcal{B}} = \begin{bmatrix} C_1 & O \end{bmatrix}$$

where $\mathcal{B}$ is an ordered basis for $\mathbb{R}^n$, the last n_o columns of which form a basis for Σ_O. Then (A_{11}, C_1) is completely observable.

Proof.

1. Let $T \in \mathbb{R}^n$ be the (invertible) matrix consisting of the columns of $\mathcal{B}$ in the same order. Then:

$$[A]_{\mathcal{B}} = V^{-1}AV, \qquad [B]_{\mathcal{B}} = V^{-1}B$$

Now, observe the following sequence of equalities:

$$\begin{aligned}
&\operatorname{rank}\left(\begin{bmatrix} B_1 & A_{11}B_1 & \cdots & A_{11}^{n_c-1}B_1 \end{bmatrix} \right) \\
&= \operatorname{rank}\left(\begin{bmatrix} B_1 & A_{11}B_1 & \cdots & A_{11}^{n_c-1}B_1 & A_{11}^{n_c}B_1 & \cdots & A_{11}^{n-1}B_1 \end{bmatrix} \right) \\
&= \operatorname{rank}\left(\begin{bmatrix} B_1 & A_{11}B_1 & \cdots & A_{11}^{n_c-1}B_1 & A_{11}^{n_c}B_1 & \cdots & A_{11}^{n-1}B_1 \\ O & O & \cdots & O & O & \cdots & O \end{bmatrix} \right) \\
&= \operatorname{rank}\left(\begin{bmatrix} [B]_{\mathcal{B}} & [A]_{\mathcal{B}}[B]_{\mathcal{B}} & \cdots & [A]_{\mathcal{B}}^{n-1}[B]_{\mathcal{B}} \end{bmatrix} \right) \\
&= \operatorname{rank}\left(V \begin{bmatrix} [B]_{\mathcal{B}} & [A]_{\mathcal{B}}[B]_{\mathcal{B}} & \cdots & [A]_{\mathcal{B}}^{n-1}[B]_{\mathcal{B}} \end{bmatrix} \right) \\
&= \operatorname{rank}\left(\begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix} \right) \\
&= n_c
\end{aligned}$$

The first equality follows from the Cayley-Hamilton theorem; since $A_{11} \in n_c \times n_c$, the columns of $\{A_{11}^{n_c}B, \cdots, A_{11}^{n-1}B\}$ are all linear combinations of the columns of $\{B, AB, \cdots, A^{n_c-1}B\}$. Thus, adding these additional columns will not affect the (overall) column rank. The second equality follows from the fact that the row rank of a matrix remains the same if extra rows of zero row vectors are added. The second-to-last equality follows from the fact that multiplication with an invertible matrix does not change rank.

2. This portion of the proof can be demonstrated similarly, or shown by considering the adjoint system of R.

■

Corollary 5.26.

1. *Here, we use the same notation used in Part a) of the previous theorem. Then the controllability matrix pencil:*
$$\begin{bmatrix} sI - A & B \end{bmatrix}$$
*lacks full row rank if and only if $s \in \sigma(A_{22})$. For this reason, the eigenvalues in A_{22} are thus called the **uncontrollable modes** of the system.*

2. *Here, we use the same notation used in Part b) of the previous theorem. Then the observability matrix pencil:*
$$\begin{bmatrix} sI - A \\ C \end{bmatrix}$$
*lacks full column rank if and only if $s \in \sigma(A_{22})$. For this reason, t The eigenvalues in A_{22} are thus called the **unobservable modes** of the system.*

Proof.

1. By retracing the proof of the above theorem, we have:
$$\begin{aligned} \operatorname{rank}\big(\begin{bmatrix} sI - A & B \end{bmatrix} \big) &= \operatorname{rank}\big(V \begin{bmatrix} sI - [A]_{\mathcal{B}} & [B]_{\mathcal{B}} \end{bmatrix} \big) \\ &= \operatorname{rank}\big(\begin{bmatrix} sI - [A]_{\mathcal{B}} & [B]_{\mathcal{B}} \end{bmatrix} \big) \\ &= \operatorname{rank}\left(\begin{bmatrix} sI - A_{11} & -A_{12} & B_1 \\ O & sI - A_{22} & O \end{bmatrix} \right) \end{aligned}$$
Since (A_{11}, B_1) is completely controllable, $\begin{bmatrix} sI - A_{11} & B_1 \end{bmatrix}$ has full row rank for each $s \in \mathbb{C}$, and thus so does $\begin{bmatrix} sI - A_{11} & -A_{12} & B_1 \end{bmatrix}$, since adding additional elements to each row of $\begin{bmatrix} sI - A_{11} & B_1 \end{bmatrix}$ does not change their linear independence (so long as these extra elements are added in the same relative positions, which, in this case, they are). Thus, $\begin{bmatrix} sI - A & B \end{bmatrix}$ loses row rank if and only if $sI - A_{22}$ loses row rank, i.e. if and only if $s \in \sigma(A_{22})$.

2. This portion of the proof can be demonstrated similarly, or shown by considering the adjoint system of R.

■

5.6 Lecture 20

Definition 5.27 (**Stabilizable**). *The linear time-invariant system (A, B) is called **stabilizable** if all uncontrollable modes are already stable, i.e. if all unstable modes are controllable. Mathematically, we have:*

$$\operatorname{rank}\big(\begin{bmatrix} sI - A & B \end{bmatrix}\big)$$

for each $s \in \sigma(A) \cap \overline{\mathbb{C}^+}$.

Definition 5.28 (**Detectable**). *The linear time-invariant system (A, C) is called **detectable** if all unobservable modes are already stable, i.e. if all unstable modes are observable. Mathematically, we have:*

$$\operatorname{rank}\left(\begin{bmatrix} sI - A \\ C \end{bmatrix}\right)$$

for each $s \in \sigma(A) \cap \overline{\mathbb{C}^+}$.

Definition 5.29. *Two systems R and $\overline{R}$, with states represented by x and $\overline{x}$, respectively, are said to be **equivalent** if there exists some:*

$$R: \quad \begin{cases} \dot{\overline{x}} = T^{-1}ATx + T^{-1}Bu, \\ y = CT\overline{x} + Du, \end{cases},$$

$$R: \quad \begin{cases} \dot{\overline{x}} = A\overline{x} + Bu, \\ y = Cx + Du, \end{cases}.$$

The systems R and R' can be considered the same system, subject to the change of basis $x = T\overline{x}$ in the state space.

Proposition 5.30 (**Eigenvalue Placement by State Feedback**). *Consider the system $R : (A, B)$, given by:*

$$\dot{x} = Ax + bu$$

where $A \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^n$ and $uin\mathbb{R}$, i.e. this is a single-input-single output (SISO) system. Then the following statements are equivalent:

1. *(A, b) is completely controllable.*
2. *There exists a matrix representation of A in the controllable canonical form, i.e. there exists some invertible $T \in \mathbb{R}^{n\times n}$ such that:*

$$\tilde{A} = T^{-1}AT = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -\alpha_n & -\alpha_{n-1} & -\alpha_{n-2} & \cdots & -\alpha_1 \end{bmatrix}, \qquad \tilde{b} = T^{-1}b = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix},$$

where $\alpha_1, \alpha_2, \cdots, \alpha_n$ are coefficients that appear in the characteristic polynomial of A and $\tilde{A}$, i.e.:

$$\chi_A(s) = s^n + \alpha_1 s^{n-1} + \cdots + \alpha_{n-1} s + \alpha_n$$

Proof.

" $\Rightarrow$ " : Suppose (A, b) is completely controllable. We will demonstrate the existence of T by implicit construction, i.e. we will explore the form that T must satisfy, in terms of A and b, if it exists.

From the controllable canonical form of $\tilde{A}$, and $\tilde{A} = T^{-1}AT$, $b = Te_n$, we have:

$$\begin{cases} \tilde{A}e_n &= e_{n-1} - \alpha_1 e_n, \\ \tilde{A}e_{n-1} &= e_{n-2} - \alpha_2 e_n, \\ &\vdots \\ \tilde{A}e_2 &= e_1 - \alpha_{n-1} e_n, \\ T\tilde{b} &= b \end{cases} \Rightarrow \begin{cases} Ab &= Te_{n-1} - \alpha_1 b, \\ ATe_{n-1} &= Te_{n-2} - \alpha_2 b, \\ &\vdots \\ ATe_2 &= Te_1 - \alpha_{n-1} b, \\ Te_n &= b \end{cases}$$

$$\Rightarrow \begin{cases} Te_n &= b, \\ Te_{n-1} &= (A + \alpha_1 I)b, \\ Te_{n-2} &= ATe_{n-1} + \alpha_2 b = A^2 b + \alpha_1 Ab + \alpha_2 b, \\ &\vdots \\ Te_1 &= ATe_2 + \alpha_{n-1} b = A^{n-1} b + \alpha_1 A^{n-2} b + \cdots + \alpha_{n-1} b, \end{cases}$$

$$\Rightarrow T = \underbrace{\begin{bmatrix} b & Ab & \cdots & A^{n-1}b \end{bmatrix}}_{\equiv \Sigma_C} \begin{bmatrix} \alpha_{n-1} & \alpha_{n-2} & \cdots & \alpha_2 & \alpha_1 & 1 \\ \alpha_{n-2} & \alpha_{n-3} & \cdots & \alpha_1 & 1 & 0 \\ \alpha_{n-3} & \alpha_{n-4} & \cdots & 1 & 0 & 0 \\ \vdots & \vdots & \iddots & \vdots & \vdots & \vdots \\ \alpha_1 & 1 & \cdots & 0 & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 & 0 \end{bmatrix}$$

Since (A, B) is completely controllable, Σ_C is invertible, and thus if we define T to be as shown above, it must also be invertible (the other matrix is upper left diagonal, and is thus always invertible).

" $\Leftarrow$ " : Conversely, suppose we have:

$$\tilde{A} = T^{-1}AT = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -\alpha_n & -\alpha_{n-1} & -\alpha_{n-2} & \cdots & -\alpha_1 \end{bmatrix}, \qquad \tilde{b} = T^{-1}b = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}.$$

Then we find that:

$$\tilde{b} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 1 \end{bmatrix} \qquad \tilde{A}\tilde{b} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ -\alpha_1 \end{bmatrix}, \qquad \tilde{A}^2\tilde{b} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ -\alpha_1 \\ \alpha_2 + \alpha_1^2 \end{bmatrix}$$

In short, for each $i = 0, \cdots, n-1$, the first $n-i-1$ elements of the vector $\tilde{A}^i\tilde{b}$ are 0, and the $(n-i)$-th element is 1. It follows that:

$$\Sigma_C = \begin{bmatrix} \tilde{b} & \tilde{A}\tilde{b} & \cdots & \tilde{A}^{n-1}\tilde{b} \end{bmatrix}$$

is lower right triangular with 1 as its $(i, n-i)$-th entry, for each i; thus, it has full rank, so (A, b) is controllable.

Finally, notice that since A and $\tilde{A}$ are related via a similarity transform, we have:

$$\chi_A(s) = \chi_{\tilde{A}}(s) = s^n + \alpha_1 s^{n-1} + \cdots + \alpha_{n-1}s + \alpha_n,$$

as can be verified via induction on the matrix blocks composing $\tilde{A}$.

■

Theorem 5.31. *Let (A, b) be completely controllable, where $A \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$, and let $\pi(s)$ be any monic polynomial of degree n with real coefficients. Then there exists a unique feedback $f^T \in \mathbb{R}^{1\times n}$ such that:*

$$\chi_{A+bf^T}(s) = \pi(s)$$

Moreover, f is given by:

$$f^T = -e_n^T \Sigma_C^{-1} \pi(A)$$

Proof. Let $\pi_1, \cdots, \pi_n \in \mathbb{R}$ be given such that:

$$\pi(s) = s^n + \pi_1 s^{n-1} + \cdots + \pi_{n-1}s + \pi_n.$$

Again, we demonstrate the existence of f^T via explicit construction. First, notice that:

$$\chi_{A+bf^T}(s) = \chi_{T^{-1}(A+bf^T)T} = \chi_{\tilde{A}+\tilde{b}\tilde{f}^T}$$

where, in addition to $\tilde{A} = T^{-1}AT, \tilde{b} = T^{-1}b$, as defined in the previous theorem, we have also defined:

$$\tilde{f}^T \equiv f^T T = \begin{bmatrix} f_n & \cdots & f_1 \end{bmatrix}$$

We wish to determine the identity of $f_1, \cdots, f_n$, as doing so would determine $\tilde{f}^T$ and thus (using the formula for T in the above theorem) f. To that end, observe that:

$$\because \tilde{A} + \tilde{b}\tilde{f}^T = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -\alpha_n + f_n & -\alpha_{n-1} + f_{n-1} & -\alpha_{n-2} + f_{n-2} & \cdots & -\alpha_1 + f_1 \end{bmatrix}$$

It thus follows that:

$$\chi_{A+bf^T} = s^n + (\alpha_1 - f_1)s^{n-1} + \cdots + (\alpha_{n-1} - f_{n-2})s + (\alpha_n - f_n)$$

Since we want:

$$\chi_{A+bf^T} = \pi(s) = s^n + \pi_1 s^{n-1} + \cdots + \pi_{n-1}s + \pi_n$$

We thus must take $f_i = \alpha_i - \pi_i$ for each $i = 1, \cdots, n$, i.e. we want:

$$\tilde{f}^T = f^T T = \begin{bmatrix} \alpha_n - \pi_n & \cdots & \alpha_1 - \pi_1 \end{bmatrix}$$

Since T is invertible, we have thus found an explicit, achievable (from a design point of view) formula for f^T.

■

Remark. In particular, the state feedback that achieves the $\chi_{A+bf^T}(s) = \pi(s)$ is:

$$f^T = -\begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix} \begin{bmatrix} b & Ab & \cdots & A^{n-1}b \end{bmatrix}^{-1} \pi(A)$$

(assuming, naturally, that (A, b) is controllable). This can be shown as follows. Observe that:

$$\begin{aligned} e_1^T \tilde{A} &= e_2^T, \\ e_1^T \tilde{A}^2 &= e_2^T \tilde{A} = e_3^T, \\ &\vdots \\ e_1^T \tilde{A}^n &= e_2^T \tilde{A}^{n-1} = \cdots = e_{n-1}^T \tilde{A} = \begin{bmatrix} -\alpha_n & \cdots & -\alpha_1 \end{bmatrix} \end{aligned}$$

Thus, we want:

$$\begin{aligned} \tilde{f}^T = f^T T &= \begin{bmatrix} \alpha_n - \pi_n & \cdots & \alpha_1 - \pi_1 \end{bmatrix} \\ &= -\pi_n e_1^T - \pi_{n-1} e_2^T - \cdots - \pi_2 e_{n-1}^T - \pi_1 e_n^T - \begin{bmatrix} -\alpha_n & \cdots & -\alpha_1 \end{bmatrix} \\ &= -\pi_n e_1^T - \pi_{n-1} e_1^T \tilde{A} - \cdots - \pi_2 e_1^T \tilde{A}^{n-2} - \pi_1 e_1^T \tilde{A}^{n-1} - e_1^T \tilde{A}^n \\ &= -e_1^T \left[\pi_n + \pi_{n-1}\tilde{A} + \cdots + \pi_1 \tilde{A}^{n-1} + \tilde{A}^n \right] \\ &= -e_1^T \pi(\tilde{A}) \end{aligned}$$

Multiplying on the right by T^{-1}, we have:

$$\begin{aligned} f^T &= -e_1^T \pi(\tilde{A}) T^{-1} = -e_1^T \pi(T^{-1}AT)T^{-1} = -e_1^T T^{-1} \pi(A) \\ &= -e_1^T \begin{bmatrix} \alpha_{n-1} & \alpha_{n-2} & \cdots & \alpha_2 & \alpha_1 & 1 \\ \alpha_{n-2} & \alpha_{n-3} & \cdots & \alpha_1 & 1 & 0 \\ \alpha_{n-3} & \alpha_{n-4} & \cdots & 1 & 0 & 0 \\ \vdots & \vdots & \iddots & \vdots & \vdots & \vdots \\ \alpha_1 & 1 & \cdots & 0 & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 & 0 \end{bmatrix}^{-1} \begin{bmatrix} b & Ab & \cdots & A^{n-1}b \end{bmatrix}^{-1} \pi(A) \\ &= -e_n^T \Sigma_C^{-1} \pi(A) \end{aligned}$$

since the first row of the inverse of the left upper triangular matrix whose elements consist of $1, \alpha_1, \cdots, \alpha_{n-1}$ is e_n^T.

Remark. We considered *negative* state feedback ($u = -Fx$) above; for the rest of this lecture, we will use *positive* state feedback ($u = Fx$). To avoid confusion, readers should keep this distinction in mind, though using either gives the same results for controllability and stabilizability.

Note (**Notation**). Below, we will use the notation $\Sigma_C(A, B) \equiv \begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix}$.

Theorem 5.32 (**State Feedback and Controllability**). *Consider an LTI system (A, B) with state feedback:*

$$\begin{aligned} \dot{x} &= Ax + Bu, \\ u &= Kx + v. \end{aligned}$$

Then $R\big(\Sigma_C(A, B)\big) = R\big(\Sigma_C(A+BK, B)\big)$. In particular, $(A+BK, B)$ is completely controllable if and only if (A, B) is.

Proof. First, for each $i = 1, \cdots, n-2$, the expression $(A + BK)^i B$ can be expanded as follows:

$$\begin{aligned} &(A + BK)^i B = A^i B + A^{i-1} B M_{i,i-1} + \cdots + B M_{i,0}, \\ \iff &A^i B = (A + BK)^i B + A^{i-1} B(-M_{i,i-1}) + \cdots + B(-M_{i,0}), \end{aligned}$$

for some matrices $M_{i,0}, M_{i,1}, \cdots, M_{i,i-1} \in \mathbb{R}^{n_i \times n_i}$, where we have labeled the second subscript of each M to match the corresponding power of A in each term. We thus have:

$$\begin{aligned} R\big((A + BK)^i B\big) &\subset R(A^i B) + R(A^{i-1} B) + \cdots + R(B), \\ R(A^i B) &\subset R\big((A + BK)^i B\big) + R(A^{i-1} B) + \cdots + R(B), \end{aligned}$$

Replacing i with $j = 0, \cdots, i-1$ in the second statement, and substituting the resulting relations back into the second statement itself, we have:

$$\begin{aligned} R\big((A + BK)^i B\big) &\subset R(A^i B) + R(A^{i-1} B) + \cdots + R(B), \\ R(A^i B) &\subset R\big((A + BK)^i B\big) + R\big((A + BK)^{i-1} B\big) + \cdots + R(B), \end{aligned}$$

Thus:

$$\begin{aligned} R(\Sigma_C(A+BK,B) &= R\big((A+BK)^{n-1}B\big) + \cdots + R\big((A+BK)B\big) + R(B) \\ &= R(A^{n-1}B) + \cdots + R(AB) + R(B) \\ &= R(\Sigma_C(A,B)), \end{aligned}$$

concluding the proof. ■

Remark. Alternatively, we could have demonstrated the equivalence of the controllability of $(A+BK,B)$ with that of (A,B) (a weaker claim than $R\big(\Sigma_C(A+BK,B)\big) = R\big(\Sigma_C(A,B)\big)$) by applying the PBH test. For each $s \in \mathbb{C}$, we have:

$$\big[sI-(A+BK),B\big] = \big[sI-A,B\big]\begin{bmatrix} I_{n\times n} & 0 \\ -K & I_{n_i\times n_i}\end{bmatrix}.$$

Since the matrix $\begin{bmatrix} I_{n\times n} & 0 \\ -K & I_{n_i\times n_i}\end{bmatrix}$ is invertible, it follows that $\big[sI-(A+BK),B\big]$ is of full row rank if and only if $\big[sI-A,B\big]$ is of full row rank. The PBH test thus gives the desired result.

Theorem 5.33 (**Output Feedback and Controllability**). *Consider the LTI system with output feedback:*

$$\begin{aligned} \dot{x} &= Ax + Bu, \\ u &= Ly + v. \end{aligned}$$

The following statements hold:

1. $R\big(\Sigma_C(A+BLC,B)\big) = R\big(\Sigma_C(A,B)\big)$.
2. $N\big(\Sigma_O(A+BLC,C)\big) = N\big(\Sigma_O(A,C)\big)$.

Thus, $(A+BLC,C)$ is completely controllable or completely observable if and only if (A,B) is.

Proof.

1. This follows from Theorem 5.32; take $K = LC$.
2. We use a similar strategy as the one used to prove Theorem 5.32. By expanding $C(A+BLC)^i$, we find that:

$$\begin{aligned} C(A+BLC)^i &= CA^i + M_{i,i-1}CA^{i-1} + \cdots + M_{i,0}C \\ &\Leftrightarrow CA^i = C(A+BLC)^i + (-M_{i,i-1})CA^{i-1} + \cdots + (-M_{i,0})C \end{aligned}$$

for some $M_{i,i-1}, \cdots, M_{i,0} \in \mathbb{R}^{n_o\times n_o}$. This implies that:

$$\begin{aligned} N(CA^i) \cap N(CA^{i-1}) \cap \cdots \cap N(C) &\subset N\big(C(A+BLC)^i\big), \\ N\big(C(A+BLC)^i\big) \cap N(CA^{i-1}) \cap \cdots \cap N(C) &\subset N(CA^i). \end{aligned}$$

Replacing i with j for each $j = 0, \cdots, i-1$ in the second statement above, and substituting the resulting relations back into the second statement itself, we find that:

$$N(CA^i) \cap N(CA^{i-1}) \cap \cdots \cap N(C) \subset N\big(C(A+BLC)^i\big),$$
$$N\big(C(A+BLC)^i\big) \cap N\big(C(A+BLC)^{i-1}\big) \cap \cdots \cap N(C) \subset N(CA^i).$$

which shows that:

$$\begin{aligned} N\big(\Sigma_O(A+BLC, C)\big) &= N\big(C(A+BLC)^{n-1}\big) \cap \cdots \cap N\big(C(A+BLC)\big) \cap N(C) \\ &= N(CA^{n-1}) \cap \cdots \cap N(CA) \cap N(C) \\ &= N\big(\Sigma_O(A, C)\big), \end{aligned}$$

completing the proof.

■

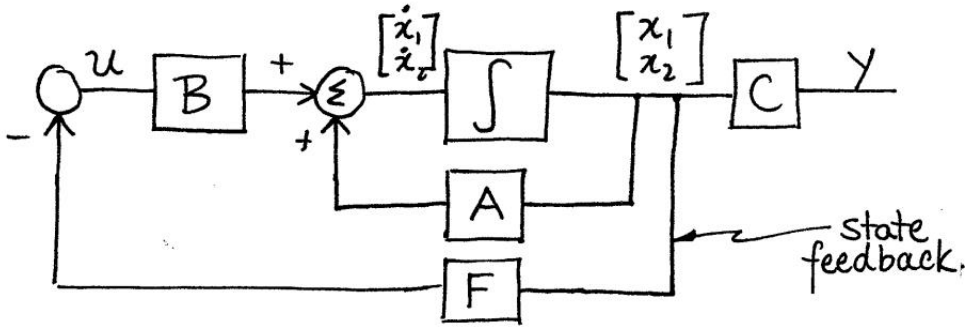
Example (**Lecture 20, pg. 7**). Consider the system:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} -1 & a \\ 3 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u$$

For which values of a are we able to place the poles of the closed loop system in *any* desired location? In particular, try to design a state feedback that relocates the poles of the system to $\lambda = -2, -3$.

Solution :

We apply state feedback, as shown in the figure below:



Let the feedback be given by $u = -Fx$, where $F = \begin{bmatrix} f_1 & f_2 \end{bmatrix}$. We have $\dot{x} = (A - BF)x$, where:

$$\begin{aligned} A - BF &= \begin{bmatrix} -1 & a \\ 3 & -2 \end{bmatrix} - \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} f_1 & f_2 \end{bmatrix} \\ &= \begin{bmatrix} -1 & a \\ 3 - f_1 & -2 - f_2 \end{bmatrix} \\ \Rightarrow \chi_{A-BF}(s) &= \det\left(\begin{bmatrix} s+1 & -a \\ f_1 - 3 & s+2+f_2 \end{bmatrix} \right) \\ &= s^2 + (3+f_2)s + \big(2 + f_2 + a(f_1 - 3)\big) \end{aligned}$$

We want the characteristic function to assume the form:

$$(s+2)(s+3) = s^2 + 5s + 6,$$

If $a \neq 0$, we can take:

$$F = \begin{bmatrix} f_1 & f_2 \end{bmatrix} = \begin{bmatrix} \frac{2}{a} + 3 & 2 \end{bmatrix}$$

Thus, the state feedback control law is:

$$\begin{aligned} u = -Fx &= \begin{bmatrix} f_1 & f_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = -f_1 x_1 - f_2 x_2 \\ &= -\left(\frac{2}{a} + 3\right) x_1 - 2x_2 \end{aligned}$$

Note that this result depends on the condition $a \neq 0$. On the other hand, if $a = 0$, we have:

$$\begin{aligned} \chi_{A-BF}(s) &= s^2 + (3 + f_2)s + (2 + f_2) \\ &= (s+1)\big(s + (2 + f_2)\big) \end{aligned}$$

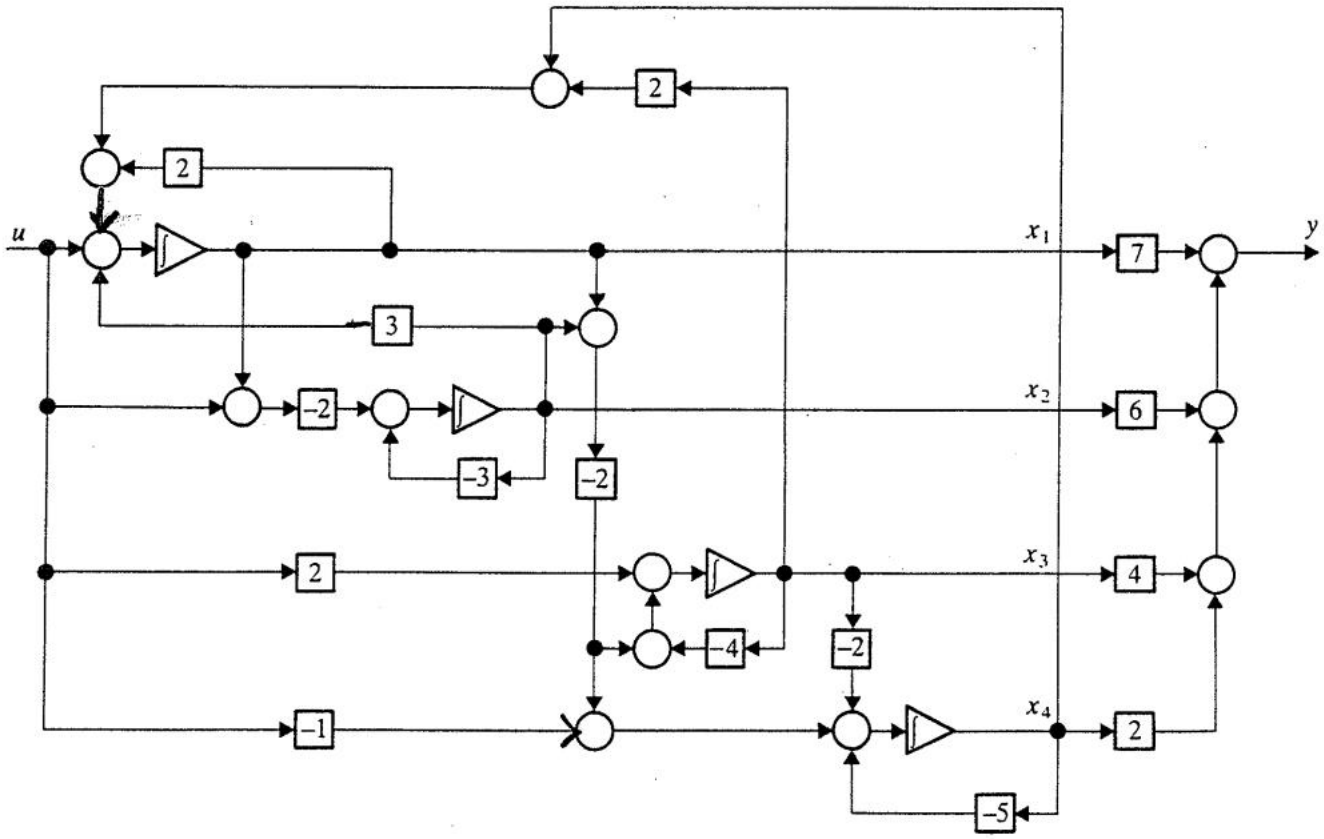
Thus, regardless of our choice of f_1, f_2, we cannot move the pole at $s = -1$. However, we can still relocate the pole at $s = -2$ to an arbitrary location, say, $s = \lambda$, by taking $f_2 = -\lambda - 2$.

In summary, if $a = 0$, the pole originally at $\lambda = -1$ is fixed, while the pole at $\lambda = 2$ can be relocated to an arbitrary location via state feedback. On the other hand, if $a \neq 0$, then both poles can be moved to arbitrary locations.

Example (**Lecture 20, pg. 10**). Consider the following system:

$$\begin{aligned} \dot{x}_1 &= 2x_1 + 3x_2 + 2x_3 + x_4 + u, \\ \dot{x}_2 &= -2x_1 - 3x_2 - 2u, \\ \dot{x}_3 &= -2x_1 - 2x_2 - 4x_3 + 2u, \\ \dot{x}_4 &= -2x_1 - 2x_2 - 2x_3 - 5x_4 - u, \\ y &= 7x_1 + 6x_2 + 4x_3 + 2x_4, \end{aligned}$$

as shown in the figure below. Find its transfer function, and interpret the result.



Solution :

The given dynamics can be expressed in matrix form, as shown below:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{bmatrix} = \underbrace{\begin{bmatrix} 2 & 3 & 2 & 1 \\ -2 & -3 & 0 & 0 \\ -2 & -2 & -4 & 0 \\ -2 & -2 & -2 & -5 \end{bmatrix}}_{\equiv A} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \underbrace{\begin{bmatrix} 1 \\ -2 \\ 2 \\ -1 \end{bmatrix}}_{\equiv B} u,$$

$$y = \underbrace{\begin{bmatrix} 7 & 6 & 4 & 2 \end{bmatrix}}_{\equiv C} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$$

Rather than solve for the transfer function by directly taking the Laplace transform of the complicated-looking matrices above, we will first try to find a similarity transform for the system by diagonalizing A. Recall that the equivalent system representation for an LTI system $\Sigma : (A, B, C, D)$, with state transformation $\overline{x} = Px$, is $\overline{\Sigma} : (PAP^{-1}, PB, CP^{-1}, D)$. In this

case, we wish to associate P with the matrix transformation from x to $\overline{x}$, i.e.:

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

$$= \underbrace{\begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix}}_{\equiv P^{-1}} \underbrace{\begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & -2 & 0 & 0 \\ 0 & 0 & -3 & 0 \\ 0 & 0 & 0 & -4 \end{bmatrix}}_{\equiv \overline{A}} \underbrace{\begin{bmatrix} 4 & 3 & 2 & 1 \\ 3 & 3 & 2 & 1 \\ 2 & 2 & 2 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}}_{\equiv P}$$

$$\Rightarrow \overline{A} = PAP^{-1} = \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & -2 & 0 & 0 \\ 0 & 0 & -3 & 0 \\ 0 & 0 & 0 & -4 \end{bmatrix},$$

$$\overline{B} = PB = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix},$$

$$\overline{C} = CP^{-1} = \begin{bmatrix} 1 & 1 & 0 & 0 \end{bmatrix}$$

Since equivalent systems have the same transfer functions, we have:

$$H(s) = C(sI - \overline{A})^{-1}B = \frac{1}{s+1}$$

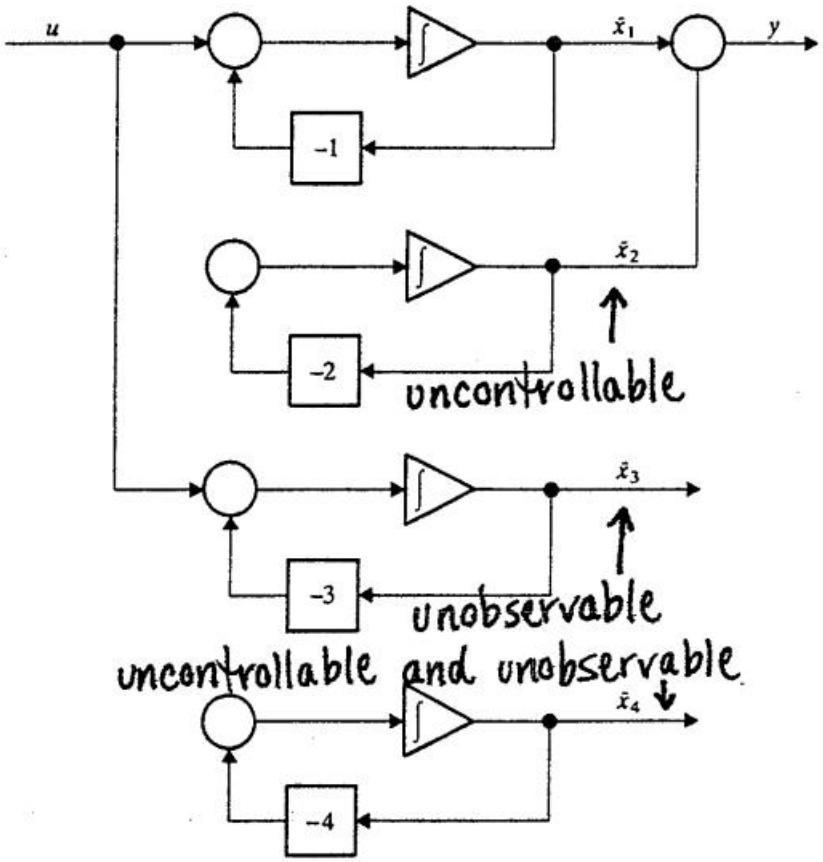
In other words, since $\dot{\overline{x}} = \overline{A}\overline{x} + \overline{B}u$, $y = \overline{C}\overline{x}$, we have:

$$\begin{aligned} \dot{\overline{x_1}} &= -\overline{x_1} + u, \\ \dot{\overline{x}}_2 &= -2\overline{x}_2, \\ \dot{\overline{x}}_3 &= -3\overline{x}_3 + u, \\ \dot{\overline{x}}_4 &= -4\overline{x}_4, \\ y &= \overline{x_1} + \overline{x}_2 \end{aligned}$$

We interpret the above results as follows.

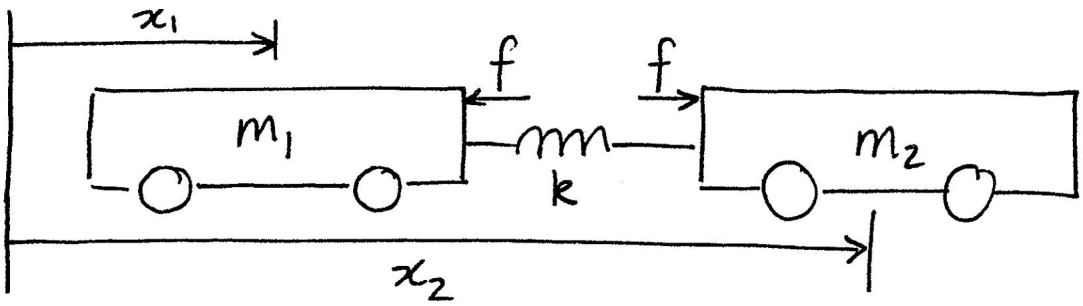
$\overline{x_1}$:	Affected by the input, Visible in the output,
$\overline{x}_2$:	Not affected by the input, Visible in the output,
$\overline{x}_3$:	Affected by the input, Not visible in the output,
$\overline{x}_4$:	Not affected by the input, Not visible in the output,

The figure below demonstrates the relationship between $\overline{x_1}, \overline{x}_2, \overline{x}_3$, and $\overline{x}_4$:



In some cases, the controllability of a system can be explicitly associated with, or interpreted by, physical characteristics of the system we are describing. Consider, for instance, the next example.

Example (**Lecture 20, pg. 13**). Consider the following system, which is *physically uncontrollable*. This is because the only forces and torques ("inputs") are internal to the system, and Newton's Third Law—every action has an equal and opposite reaction—implies that the center of mass of a closed system cannot be changed by internal forces or torques.



Let x_1, x_2 be the center of masses of m_1, m_2, respectively, as shown above, and let $x_3 \equiv \dot{x}_1, x_4 \equiv \dot{x}_2$ denote their respective velocities. The dynamics of the system can then be modeled as:

$$\begin{aligned}
\dot{x}_1 &= x_3, \\
\dot{x}_2 &= x_4, \\
\dot{x}_3 &= -\frac{k}{m_1}(x_1 - x_2) - \frac{f}{m_1}, \\
\dot{x}_4 &= -\frac{k}{m_2}(x_2 - x_1) + \frac{f}{m_2}.
\end{aligned}$$

Check the controllability of the system, and interpret your results.

Solution:

Rewriting the above equations in matrix form, we have:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -\frac{k}{m_1} & \frac{k}{m_1} & 0 & 0 \\ \frac{k}{m_2} & -\frac{k}{m_1} & 0 & 0 \end{bmatrix}}_{\equiv A} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \underbrace{\begin{bmatrix} 0 \\ 0 \\ -\frac{1}{m_1} \\ \frac{1}{m_2} \end{bmatrix}}_{\equiv B} u.$$

We can check the controllability of the system by finding the rank of the controllability matrix:

$$\begin{aligned}
\Sigma_C &\equiv \begin{bmatrix} B & AB & A^2B & A^3B \end{bmatrix} \\
&= \begin{bmatrix} 0 & -\frac{1}{m_1} & 0 & \frac{k}{m_1}\left(\frac{1}{m_1} + \frac{1}{m_2}\right) \\ 0 & \frac{1}{m_2} & 0 & -\frac{k}{m_2}\left(\frac{1}{m_1} + \frac{1}{m_2}\right) \\ -\frac{1}{m_1} & 0 & \frac{k}{m_1}\left(\frac{1}{m_1} + \frac{1}{m_2}\right) & 0 \\ \frac{1}{m_2} & 0 & -\frac{k}{m_2}\left(\frac{1}{m_1} + \frac{1}{m_2}\right) & 0 \end{bmatrix}
\end{aligned}$$

Since the second and fourth rows are $-m_1/m_2$ times the first and third rows, respectively, Σ_C is of rank 2, and thus lacks full row rank. The system is thus uncontrollable.

We can interpret this result by, once again, applying a similarity transform to the original dynamics. This time, we wish to transform the system into an equivalent representation whose first two states $\overline{x_1}, \overline{x}_2$, measure the *center of mass* of the system:

$$\overline{x_1} = \frac{m_1 x_1 + m_2 x_2}{m_1 + m_2}$$

and the displacement between the two masses:

$$\overline{x}_2 = x_1 - x_2$$

Analogous to our original system, we will define $\overline{x}_3, \overline{x}_4$, to be the rate of change of $\overline{x_1}, \overline{x}_2$. Mathematically, this requires us to consider the transformation:

$$\begin{bmatrix} \overline{x_1} \\ \overline{x}_2 \\ \overline{x}_3 \\ \overline{x}_4 \end{bmatrix} = \underbrace{\begin{bmatrix} \frac{m_1}{m_1+m_2} & \frac{m_2}{m_1+m_2} & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & \frac{m_1}{m_1+m_2} & \frac{m_2}{m_1+m_2} \\ 0 & 0 & 1 & -1 \end{bmatrix}}_{\equiv P} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$$

As stated in the above example, the equivalent system representation for an LTI system $\Sigma : (A, B, C, D)$, with state transformation $\overline{x} = Px$, is $\overline{\Sigma} : (PAP^{-1}, PB, CP^{-1}, D)$. After some algebra, we have:

$$\overline{A} = PAP^{-1} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -k\left(\frac{1}{m_1} + \frac{1}{m_2}\right) \end{bmatrix},$$

$$\overline{B} = PB = \begin{bmatrix} 0 \\ 0 \\ 0 \\ -\left(\frac{1}{m_1} + \frac{1}{m_2}\right) \end{bmatrix}$$

Notice that only the fourth element in $\overline{B}$ is nonzero; this implies that the internal forces, which constitute the only input to this system, can only affect:

$$\dot{x}_4 = x_3 - x_4 = \frac{d}{dt}(x_1 - x_2)$$

That is, although the internal forces will change the relative positions of x_1 and x_2, the center of mass of the entire system, x_1, remains completely unaffected. To control x_1, an external force is needed.

Example (**Lecture 20, pg. 16**). Consider the system:

$$\dot{x} = \underbrace{\begin{bmatrix} -1 & 0 \\ 0 & -3 \end{bmatrix}}_{\equiv A} + \underbrace{\begin{bmatrix} 1 \\ 1 \end{bmatrix}}_{\equiv B} u$$

1. Is the given system controllable?
2. If the given system is controllable, determine the *negative* state feedback gain matrix $F = \begin{bmatrix} f_1 & f_2 \end{bmatrix}$ that relocates the poles to $s = -2, -2$.

Solution :

1. The controllability matrix of the given system is:

$$\Sigma_C = \begin{bmatrix} B & AB \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 1 & -3 \end{bmatrix},$$

which has full row rank. Thus, the system is completely controllable.

2. We have:

$$\begin{aligned} A - bf^T &= \begin{bmatrix} -1 - f_1 & -f_2 \\ -f_1 & -3 - f_2 \end{bmatrix} \\ \Rightarrow \chi_{A-bf^T}(s) &= \det\left(\begin{bmatrix} s + (1 + f_1) & f_2 \\ f_1 & s + (3 + f_2) \end{bmatrix} \right) \\ &= s^2 + (4 + f_1 + f_2) + (3f_1 + f_2 + 3) \end{aligned}$$

We want the characteristic function to be:

$$(s + 2)^2 = s^2 + 4s + 4$$

Thus, we should take f_1, f_2 such that:

$$\begin{aligned} f_1 + f_2 &= 0 \\ 3f_1 + f_2 &= 1, \end{aligned}$$

i.e. $F = \begin{bmatrix} f_1 & f_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \end{bmatrix}$.

Remark. Part b) of the above problem can also be directly solved by transforming A, B to their control canonical form. As verified in Theorem 5.31, we have: (Note that the sign of F is inverted, because the theorem considers positive feedback, while here we consider negative feedback)

$$\begin{aligned} F &= e_2^T \Sigma_C^{-1} \pi(A) \\ &= \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 1 & -3 \end{bmatrix}^{-1} (A^2 + 4A + 4I) \\ &= -\frac{1}{2} \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} -3 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \end{bmatrix} \end{aligned}$$

Example (**Lecture 20, pg. 20**). An approximate linear model of the longitudinal dynamics of certain aircraft, for a particular set of conditions, has the linearized state and control vectors:

$$x = \begin{bmatrix} p \\ r \\ \beta \\ \phi \end{bmatrix}, \qquad u = \begin{bmatrix} \delta \\ \mu \end{bmatrix}$$

where the variables given above have the following physical interpretations:

States: p — incremental roll rate,
r — incremental yaw rate,
β — incremental sideslip angle,
ϕ — incremental roll angle
Inputs: δ_a — aileron angle
δ_r — rudder angle

The state space equation for this model is $\dot{x} = Ax + Bu$, where:

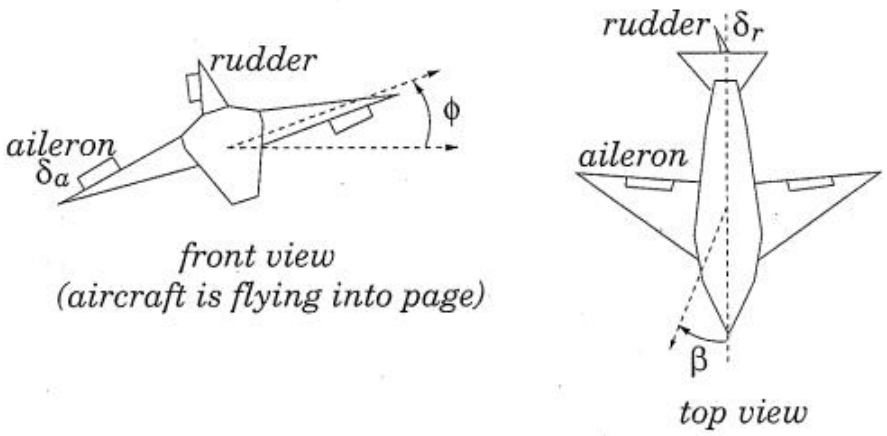
$$A = \begin{bmatrix} -10 & 0 & -10 & 0 \\ 0 & -0.7 & 9 & 0 \\ 0 & -1 & -0.7 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} 20 & 2.8 \\ 0 & -3.13 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

1. Suppose a malfunction prevents manipulation of the input δ_r. Is it possible to completely control the aircraft using only δ_a?

2. If you had your choice of only one of the following sensors, which would you use? Explain.

   - A rate gyro which measures the roll rate p.
   - A bank indicator which measures ϕ.

   A figure for the aircraft is provided below:



Solution :

1. We can analyze the controllability of the given system in the case where we are only allowed to control δ_a by considering the controllability matrices for the system in this case.

   Let B_1, B_2 denote the first and second columns of B, respectively. Then the controllability matrices under the assumption that we can control on only δ_a would be:

$$\Sigma_{C1} = \begin{bmatrix} B_1 & AB_1 & A^2B_1 & A^3B_1 \end{bmatrix} = \begin{bmatrix} 20 & -200 & 2000 & -20000 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 20 & -200 & 2000 \end{bmatrix},$$

   respectively. Since Σ_{C1} has rank 2, and thus lacks full row rank, the system is uncontrollable.

2. Similarly, we can analyze the observability of the given system in the case where we are only allowed to observe q or θ by considering the observability matrices for the system under the above two cases.

   First, note that "only observing p" and "only observing θ" correspond to the following matrices for C:

$$C_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix},$$
$$C_2 = \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix},$$

   respectively. In this case, the observability matrices under the assumption that we can control on only δ and only μ are:

$$\Sigma_{C1} = \begin{bmatrix} C_1 \\ C_1A \\ C_1A^2 \\ C_1A^3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -10 & 0 & -10 & 0 \\ 100 & 10 & 107 & 0 \\ -1000 & -114 & -984.9 & 0 \end{bmatrix},$$
$$\Sigma_{C2} = \begin{bmatrix} C_2 \\ C_2A \\ C_2A^2 \\ C_2A^3 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ -10 & 0 & -10 & 0 \\ 100 & 10 & 107 & 0 \end{bmatrix}$$

   Since Σ_{O_1} lacks full row rank (it has rank 3), while Σ_{O_2} has full row rank, we conclude that only having a bank indicator which measures ϕ is preferable to only having a rate gyro which measures the roll rate p.

Remark. The above example, taken from Bryson's text on aircraft dynamics, describes the lateral dynamics of a conventional take-off and landing (CTOL) aircraft, which simply means that the aircraft takes off and lands on a runway (as opposed to vertical take-off and landing aircrafts). These aircraft are designed such that the longitudinal dynamics (e.g. pitch, up-and-down motion) with its lateral motion. For instance, BOEING has designed such aircraft such

that the lateral and longitudinal dynamics can be controlled independently. This is not the case for sophisticated military aircraft.

The roll rate p in this example is the rate at which the plane rotates about the axis passing through its front (nose) and end (tail). The yaw rate r, meanwhile, describes rotation about a vertical axis passing through the midsection of the plane, i.e. "turning side to side." The sideslip angle β describes the difference between the direction in which the aircraft is pointing and the direction in which the aircraft is currently headed. The aileron angle δ_a is associated with the flaps at the aircraft's tail that move up and down, while the rudder angle δ_r is associated with the flaps at the aircraft's tail that move up and down. The roll angle (or bank angle) ϕ is the angle associated with, and changing at a rate equal to, the roll rate p. In particular, the roll rate p and roll angle ϕ are of interest because they provide the first indication that an aircraft will very soon be turning. This is attributed to the actual coupling between the longitudinal and lateral movement of the aircraft, a coupling ignored in the aforementioned model in which the lateral and longitudinal dynamics were assumed to be completely independent.

More details can be found in Professor Claire Tomlin's Video Lecture 31 (30:30—41:22), in which she explains the terminology and concepts associated with this exercise.

5.7 Lectures 18, 19, 20 Discussion

Example (**Discussion 12, Problem 1**). Consider the LTI system given by:

$$\dot{x} = \begin{bmatrix} 0 & 1 \\ -1 & -1 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u,$$
$$y = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} x$$

1. Is the system controllable? Is it observable?
2. Can the closed loop poles of the system be placed at $\lambda_1 = -2, \lambda_2 = -2$ using output feedback alone?

   Now, consider the same plant with an additional state measurement state such that:

   $$y = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} x$$

3. Is the system still controllable and observable?
4. Can the closed loop poles of the system be placed at $\lambda_1 = -2, \lambda_2 = -2$?
5. Explain how the closed loop poles of the system could be placed at $\lambda_1 = -2, \lambda_2 = -2$ using only a single sensor, i.e., using only a one-dimensional output.

Solution :

1. The controllability and observability matrices are, respectively:

   $$\Sigma_C = \begin{bmatrix} B & AB \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & -1 \end{bmatrix},$$
   $$\Sigma_O = \begin{bmatrix} C \\ CA \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$$

   Since Σ_C has full row rank and Σ_O has full column rank, the system is both controllable and observable.

2. This problem can be solved by considering the properties that must be satisfied by an output feedback that places the poles of the resulting closed-loop system at $\lambda_1 = -2, \lambda_2 = -2$, if it exists.

   Formally, consider an arbitrary output feedback $u = Fy$, where $F = \begin{bmatrix} f_1 & f_2 \end{bmatrix} \in \mathbb{R}^{1\times 2}$. The dynamics of the closed-loop system is thus:

   $$\dot{x} = Ax + Bu = Ax + B(Fy) = (A + BFC)y,$$

where:

$$\begin{aligned} A + BFC &= \begin{bmatrix} 0 & 1 \\ -1 & -1 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} f_1 & f_2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} 0 & 1 \\ -1 + f_1 & -1 \end{bmatrix}, \\ \chi_{A+BFC}(s) &= s(s+1) + (f_1 + 1) \\ &= s^2 + s + (f_1 + 1). \end{aligned}$$

However, we want $\lambda_1 = -2, \lambda_2 = -2$, or equivalently, the characteristic equation of the output feedback loop must be $s^2 + 4s + 4$. Clearly, this cannot be achieved with any choice of f_1, since the coefficient of s in $\chi_{A+BFC}(s)$ remains 1 regardless of our choice.

3. The given change in C does not affect Σ_C, but it does affect Σ_O, which becomes:

$$\Sigma_O = \begin{bmatrix} C \\ CA \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & -1 \end{bmatrix}$$

Since Σ_C still have full row rank and Σ_O has full column rank, the system is still both controllable and observable.

4. Unlike the case in sub-problem 2, we can place the poles at $\lambda_1 = -2, \lambda_2 = -2$ via ouptut feedback. This is because the controllability of Σ_A implies that state feedback can be used to place the resulting closed loop poles anywhere in the complex plane. Now, since $C = I_2$, the output y is in fact identically equal to the state x; thus, for this problem, output and state feedback are synonymous.

   The particular state feedback required to relocate the poles can be found in a manner similar to that shown in the above sub-problems. Suppose $F = \begin{bmatrix} f_1 & f_2 \end{bmatrix}$ is the desired output feedback. Since the resulting system is $\dot{x} = (A + BFC)x$, where:

$$\begin{aligned} A + BFC &= \begin{bmatrix} 0 & 1 \\ -1 & -1 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} f_1 & f_2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 0 & 1 \\ -1 + f_1 & -1 + f_2 \end{bmatrix}, \\ \chi_{A+BFC}(s) &= s(s - f_2 + 1) + (-f_1 + 1) \\ &= s^2 + s(-f_2 + 1) + (-f_1 + 1). \end{aligned}$$

   we require that $F = \begin{bmatrix} f_1 & f_2 \end{bmatrix} = \begin{bmatrix} -3 & -3 \end{bmatrix}$ to allow $\chi_{A+BFC}(s) = s^2 + 4s + 4 = (s+2)^2$, a result equivalent to pole placement at $\lambda_1 = -2, \lambda_2 = -2$.

5. If the output is one-dimensional, then the matrix C, which maps states to outputs, must have dimensions 1×2, while the output feedback matrix, which maps outputs to inputs,

must be scalar. Thus, in general, we have:

$$C = \begin{bmatrix} c_1 & c_2 \end{bmatrix},$$
$$F = \begin{bmatrix} f \end{bmatrix}$$

for some c_1, c_2, f. This problem essentially asks us to find a suitable combination of c_1, c_2, f such that the characteristic function of the resulting closed-loop system is $\chi_{A+BFC}(s) = (s+2)^2 = s^2 + 4s + 4$. To that end, let us calculate $A + BFC$:

$$\begin{aligned} A + BFC &= \begin{bmatrix} 0 & 1 \\ -1 & -1 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} f \end{bmatrix} \begin{bmatrix} c_1 & c_2 \end{bmatrix} \\ &= \begin{bmatrix} 0 & 1 \\ -1 + fc_1 & -1 + fc_2 \end{bmatrix}, \\ \chi_{A+BFC}(s) &= s(s - fc_2 + 1) + (-fc_1 + 1) \\ &= s^2 + s(-fc_2 + 1) + (-fc_1 + 1). \end{aligned}$$

Thus, we need to choose c_1, c_2, f such that $fc_1 = fc_2 = -3$. We can choose, for instance, $c_1 = 1, c_2 = 1, f = -3$.

Example (**Discussion 12, Problem 2**). An approximate linear model of the longitudinal dynamics of certain aircraft, for a particular set of conditions, has the linearized state and control vectors:

$$x = \begin{bmatrix} v \\ \alpha \\ \theta \\ q \end{bmatrix}, \qquad u = \begin{bmatrix} \delta \\ \mu \end{bmatrix}$$

where the variables given above have the following physical interpretations:

States: v — change in forward velocity,
α — change in angle of attack,
θ — change in pitch angle,
q — change in pitch rate
Inputs: δ — deflection of the elevators
μ — throttle position

The state space equation for this model is $\dot{x} = Ax + Bu$, where:

$$A = \begin{bmatrix} -0.045 & 0.036 & -32 & -2 \\ -0.4 & -3 & -0.3 & 250 \\ 0 & 0 & 0 & 1 \\ 0.002 & -0.04 & 0.001 & -3.2 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 & 0.1 \\ -03 & 0 \\ 0 & 0 \\ -10 & 0 \end{bmatrix}$$

1. Suppose a malfunction prevents manipulation of the input δ. Is it possible to completely control the aircraft using only μ? What if only δ is available?

2. If you had your choice of only one of the following sensors, which would you use? Would it make a difference? Explain.

   - A rate gyro which measures the pitch rate q.
   - A pitch indicator which measures θ.

Solution :

1. We can analyze the controllability of the given system in the case where we are only allowed to control μ or δ by considering the controllability matrices for the system under the above two cases.

   Let B_1, B_2 denote the first and second columns of B, respectively. Then the controllability matrices under the assumption that we can control on only δ and only μ are:

$$\Sigma_{C1} = \begin{bmatrix} B_1 & AB_1 & A^2B_1 & A^3B_1 \end{bmatrix} = \begin{bmatrix} 0 & 18.92 & 166 & -491.33 \\ -30 & -2410 & 15525 & -49106 \\ 0 & -10 & 33.2 & -9.81 \\ -10 & 33.2 & -9.81 & -589.25 \end{bmatrix},$$

$$\Sigma_{C2} = \begin{bmatrix} B_1 & AB_2 & A^2B_2 & A^3B_2 \end{bmatrix} = \begin{bmatrix} 0.1 & 0 & 0 & 0 \\ 0 & -0.04 & 0.1718 & -0.277 \\ 0 & 0 & 0 & 0.001 \\ 0 & 0 & 0.001 & -0.01 \end{bmatrix}$$

   respectively. Whereas Σ_{C1} is non-singular, Σ_{C2} is very close to losing rank on two of its rows. A conservative answer would be that complete control is theoretically possible with only δ or only μ, but doing so would be extremely difficult using μ.

2. Similarly, we can analyze the observability of the given system in the case where we are only allowed to observe q or θ by considering the observability matrices for the system under the above two cases.

   First, note that "only observing q" and "only observing θ" correspond to the following matrices for C:

$$C_1 = \begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix},$$
$$C_2 = \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix},$$

   respectively. In this case, the observability matrices under the assumption that we can

control on only δ and only μ are:

$$\Sigma_{C1} = \begin{bmatrix} C_1 \\ C_1A \\ C_1A^2 \\ C_1A^3 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0.002 & -0.04 & 0.001 & -3.2 \\ 0.001 & 0.248 & -0.055 & 0.237 \\ -0.099 & -0.753 & -0.379 & 61.185 \end{bmatrix},$$

$$\Sigma_{C2} = \begin{bmatrix} C_2 \\ C_2A \\ C_2A^2 \\ C_2A^3 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0.002 & -0.040 & 0.001 & -3.2 \\ 0.010 & 0.248 & -0.055 & 0.237 \end{bmatrix}$$

Notice that $C_2A = C_1$, which then implies that $C_2A^{n+1} = C_1A^n$ for each $n \in \mathbb{N}$. This implies that:

$$\begin{aligned} N(\Sigma_{O2}) &= \bigcap_{i=0}^{n-1} N(C_2A^i) = \bigcap_{i=0}^{n} N(C_2A^i) = N(C_2) \bigcap \left(\bigcap_{i=1}^{n} N(C_2A^i) \right) \\ &= N(C_2) \bigcap \left(\bigcap_{i=0}^{n-1} N(C_1A^i) \right) = N(C_2) \bigcap N(\Sigma_{O1}) \\ &\subset N(\Sigma_{O1}) \end{aligned}$$

where we have used the Cayley-Hamilton theorem to assert that $\bigcap_{i=1}^{n} N(C_2A^{i-1}) \subset N(C_2A^n)$ to justify the second equality.

In plain English, compared to the case where we only observe q (corresponding to the larger $N(\Sigma_{O1})$), we obtain a strictly smaller set of completely unobservable sets if we observe only θ (corresponding to the smaller $N(\Sigma_{O2})$). In that sense, it is better to only measure θ than to only measure q.

However, if we are only considering the observability of the two systems by asking whether Σ_{C1} and Σ_{C2} have full column rank, it makes no difference.

5.8 Lecture 21

Lemma 5.34 (Heymann Lemma). *Let (A, B) be completely controllable, with $B = \begin{bmatrix} b_1, \cdots, b_{n_i} \end{bmatrix}$, and suppose B has full rank. Then there exists a linear state feedback $u = kx + v$ such that the resulting closed-loop system*

$$\dot{x} = (A + BK)x + Bv,$$

is controllable via $b_1 v_1$, i.e. $(A + BK, b_1)$ is completely controllable.

Proof. We offer a proof of construction. Define:

$$\begin{aligned} z_1 &= b, \\ z_{i+1} &= Az_i + bv_i, \end{aligned}$$

where, for each $i \geq 1$, we choose v_i such that $\{z_1, \cdots, z_i, z_{i+1}\}$ is linearly independent.

We claim that the induction continues until $i = n$. This is because, if the induction continues up to some $i \in \{1, \cdots, n\}$, we have for each $v_i \in \mathbb{R}^{n_i}$:

$$Az_i + Bv_i \in \underbrace{\text{span}\{z_1, \cdots, z_i\}}_{\equiv M} \subset \Sigma,$$

Observe the following facts:

1. M is a subspace of Σ; this follows from its definition as the span of a set of vectors.
2. M is A-invariant, since if we take $v_i = 0$, we find $Az_i \in M$.
3. $R(B) \in M$. This is because, if not, then there exists some $v_i \in \mathbb{R}^{n_i}$ such that $Bz_i \notin M$. But then $z_{i+1} = Az_i + Bz_i \notin M$, in contradiction to our hypothesis.

By Theorem 5.22, $R(\Sigma_C)$ is the smallest A-invariant subspace of Σ that contains $R(B)$. This implies that $R(\Sigma C) \subset M$, and so $i = \text{rank}(M) \geq \text{rank}(\Sigma C) = n$, where the final equality follows from the complete controllability of (A, B). However, since M is a subspace of Σ, clearly $i \leq n$. We thus have $i = n$, as desired.

Now, since $M = \Sigma$, we find that $\{z_1, \cdots, z_n\}$ are linearly independent. Thus, fix some arbitrary $v \in \Sigma$, and define:

$$\begin{aligned} F \equiv \begin{bmatrix} z_1 & \cdots & z_{n-1} & z_n \end{bmatrix}^{-1} \begin{bmatrix} v_1 & \cdots & v_{n-1} & v \end{bmatrix} \\ \Rightarrow \begin{bmatrix} z_1 & \cdots & z_{n-1} & z_n \end{bmatrix} F = \begin{bmatrix} v_1 & \cdots & v_{n-1} & v \end{bmatrix} \end{aligned}$$

As a result, we have:

$$\begin{aligned} z_1 &= b, \\ z_2 &= Az_1 + Bv_1 = (A + BF)b, \\ z_3 &= Az_2 + Bv_2 = (A + BF)v_2 = (A + BF)^2 b, \\ &\vdots \\ z_n &= Az_{n-1} + Bv_{n-1} = (A + BF)z_{n-1} = (A + BF)^{n-1} b \end{aligned}$$

This can be rearranged as:

$$\begin{aligned} &\begin{bmatrix} b & (A+BF)b & \cdots & (A+BF)^{n-1} \end{bmatrix}, \\ =&\begin{bmatrix} z_1 & z_2 & \cdots & z_n \end{bmatrix} \end{aligned}$$

In other words, the controllability matrix pencil for $(A+BF, b)$ has rank n, i.e. full row rank. We conclude that $(A+BF, b)$ is completely controllable.

■

Remark. The above lemma tells that, if (A, B) is completely controllable and $b \in R(B)$, then we can find a suitable state feedback F that moves the eigenvalues of A to wherever we want.

Formally, given $A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times n_i}$, and any monic polynomials $\pi(s)$ with real coefficients, choose $b \in R(B)$ arbitrarily, and let v be given such that $b = Bv$. Then there exists some $F_1 \in \mathbb{R}^{n_i\times n}$ such that $(A+BF_1, b)$ is completely controllable.

Now, by the remark following Theorem 5.31, choose:

$$f_2^T = e_n^T \begin{bmatrix} b & (A+BF_1)b & \cdots & (A+BF)^{n-1}b \end{bmatrix}^{-1} \pi(A+BF_1).$$

Then, observe that:

$$\sigma(A+BF_1+bf_2^T) = \sigma(A+B(F_1+vf_2^T))$$

consist of the roots of $\pi(s)$; in other words, $F = F_1 + vf_2^T$ places the closed loop eigenvalues of the system at $\pi(s)$.

Lemma 5.35. *Given an LTI system $R : \dot{x} = Ax + Bu$, if (A, B) is not completely controllable, then for each state feedback $F \in \mathbb{R}^{n_i\times n}$, the resulting closed-loop dynamics contain all uncontrollable modes, i.e. if $\lambda \in \sigma(A)$, with $\operatorname{rank}\big(\begin{bmatrix} \lambda I - A & B \end{bmatrix}\big) < n$, then $\lambda \in \sigma(A+BF)$ for each $F \in \mathbb{R}^{n_i\times n}$.*

Proof. Suppose $\operatorname{rank}\big(\begin{bmatrix} sI - A & B \end{bmatrix}\big) < n$ for some $s \in \mathbb{C}$. Then there exists some $v \neq 0$ such that:

$$\begin{aligned} &v^T \begin{bmatrix} sI - A & B \end{bmatrix} = 0, \\ \Rightarrow &v^T \begin{bmatrix} sI - (A+BK) & B \end{bmatrix} = v^T \begin{bmatrix} sI - A & B \end{bmatrix} \begin{bmatrix} I & O \\ -K & I \end{bmatrix} = 0, \end{aligned}$$

so $\operatorname{rank}\big(\begin{bmatrix} sI - (A+BK) & B \end{bmatrix}\big) < n$ as well. This implies that $s \in \sigma(A+BK)$, since $sI - (A+BK)$ can only have rank less than n when $s \in \sigma(A+BK)$.

■

Theorem 5.36 (PBH Test for Stabilizability). *Given an LTI system (A, B), the following two statements are equivalent (the first is simply the definition of stabilizability):*

1. $\operatorname{rank}\big(\begin{bmatrix} sI - A & B \end{bmatrix}\big) = n, \qquad \forall s \in \overline{\mathbb{C}^+}.$
2. *$\exists F \in \mathbb{R}^{n_i\times n}$ such that $\sigma(A+BK) \subset \mathbb{C}^-$.*

Proof.

" $\Leftarrow$ " : Suppose by contradiction that there exists some $\lambda \in \overline{C^+}$ such that $\text{rank}(\lambda I - A, B) < n$. Then, by the above lemma, $\lambda \in \sigma(A+BF)$, contradicting the fact that $\sigma(A+BK) \in \mathbb{C}^-$.

" $\Rightarrow$ " :

Now, suppose $\text{rank}\big(\begin{bmatrix} sI - A & B \end{bmatrix}\big) = n$ for each $s \in \overline{C^+}$. The theory of Kalman decomposition, we can find an invertible matrix T whose columns consist of basis vectors formed from bases vectors of the controllable and uncontrollable subspaces, in that order. Thus, the matrix representations of A and B with respect to T are then of the form:

$$[A] = \begin{bmatrix} A_{11} & A_{12} \\ O & A_{22} \end{bmatrix}, \qquad [B] = \begin{bmatrix} B_1 \\ 0 \end{bmatrix},$$

with $A_{11} \in \mathbb{R}^{n_c \times n_c}$, $A_{12} \in \mathbb{R}^{n_c \times (n-n_c)}, A_{22} \in \mathbb{R}^{(n-n_c)\times(n-n_c)}$ and $B_1 \in \mathbb{R}^{n_c \times n_i}$, where $n_c \in \{1, \cdots, n\}$ is the dimension of the controllable subspace. By hypothesis, all of the uncontrollable modes are in the left half complex plane, so $\sigma(A_{22}) \in \mathbb{C}^-$, and (A_{11}, B_1) is completely controllable. Thus, there exists some $F_1 \in \mathbb{R}^{n_i \times n_c}$ such that $\sigma(A_{11} + B_1F_1) \in \mathbb{C}^-$. Fix some arbitrary $F_2 \in \mathbb{R}^{n_i \times (n-n_c)}$, and choose $F \equiv \begin{bmatrix} F_1 & F_2 \end{bmatrix}$. Then:

$$A + BF = \begin{bmatrix} A_{11} + B_1F_1 & A_{12} + B_1F_2 \\ O & A_{22} \end{bmatrix}$$

with $\sigma(A + BF) = \sigma(A_{11} + B_1F_1) \cup \sigma(A_{22}) \in \mathbb{C}^-$. ■

Theorem 5.37 (**PBH Test for Detectability**). *Given an LTI system (A, C), the following two statements are equivalent (the first is simply the definition of detectability):*

1. $\text{rank}\left(\begin{bmatrix} sI - A \\ C \end{bmatrix}\right) = n, \qquad \forall s \in \overline{\mathbb{C}^+}.$
2. *$\exists L \in \mathbb{R}^{n \times n_o}$ such that $\sigma(A + LC) \subset \mathbb{C}^-$.*

Proof.

" $\Leftarrow$ " : Again, the proof follows by contradiction. If there exists some $s \in \overline{\mathbb{C}^+}$ such that $\begin{bmatrix} sI - A \\ C \end{bmatrix}$ lacks full column rank, then there exists some $v \neq 0$ such that, for this $s \in \mathbb{C}^+$, and any $L \in \mathbb{R}^{n \times n_o}$:

$$\begin{bmatrix} sI - A \\ C \end{bmatrix} v = 0$$
$$\Rightarrow \begin{bmatrix} I & -L \\ O & I \end{bmatrix} \begin{bmatrix} sI - A \\ C \end{bmatrix} v = 0$$
$$\Rightarrow \begin{bmatrix} sI - (A + LC) \\ C \end{bmatrix} v = 0,$$

i.e. $\text{rank}\left(\begin{bmatrix} sI-(A+LC) \\ C \end{bmatrix}\right) < n$. This implies $s \in \sigma(A+LC)$, since $sI - A$ only have rank less than n when $s \in \sigma(A+LC)$. To summarize, there exists some $s \in \overline{\mathbb{C}^+}$ such that, for each $L \in \mathbb{R}^{n\times n_o}$, we have $s \in \sigma(A+LC)$. This is the exact opposite of the statement "$\exists L \in \mathbb{R}^{n\times n_o}$ such that $\sigma(A+LC) \subset \mathbb{C}^-$", so we are done.

" $\Rightarrow$ " : Suppose $\text{rank}\left(\begin{bmatrix} sI-A \\ C \end{bmatrix}\right) = n$ for each $s \in \overline{C^+}$. Then, by Kalman decomposition, we can find an invertible matrix T whose columns consist of basis vectors formed from bases vectors of the observable and unobservable subspaces, in that order. The matrix representations of A and C with respect to T are then of the form:

$$[A] = \begin{bmatrix} A_{11} & O \\ A_{21} & A_{22} \end{bmatrix}, \qquad [C] = \begin{bmatrix} C_1 & O \end{bmatrix},$$

where $\sigma(A_{22}) \in \mathbb{C}^-$, with $A_{11} \in \mathbb{R}^{n_{obs}\times n_{obs}}$, $A_{12} \in \mathbb{R}^{n_{obs}\times(n-n_{obs})}$, $A_{22} \in \mathbb{R}^{(n-n_{obs})\times(n-n_{obs})}$, and $C_1 \in \mathbb{R}^{n_o\times n_{obs}}$, where $n_{obs} \in \{1, \cdots, n\}$ is the dimension of the observable subspace, and n_o, as usual, is the dimension of the output. By hypothesis, all of the observable modes are in the left half complex plane, so $\sigma(A_{22}) \in \mathbb{C}^-$, and (A_{11}, C_1) is completely observable. Then there exists some $L_1 \in \mathbb{R}^{n_{obs}\times n_o}$ such that $\sigma(A_{11}+L_1C_1) \in \mathbb{C}^-$. Fix some $L_2 \in \mathbb{R}^{(n-n_{obs})\times n_o}$ arbitrarily, and define $L \equiv \begin{bmatrix} L_1 & L_2 \end{bmatrix}^T$. Then:

$$A + LC = \begin{bmatrix} A_{11}+L_1C_1 & O \\ A_{21}+L_2C_1 & A_{22} \end{bmatrix},$$

with $\sigma(A+LC) = \sigma(A_{11}+L_1C_1) \cup \sigma(A_{22}) \in \mathbb{C}^-$.

■

Theorem 5.38. *An LTI system (A, C) is detectable if and only if its adjoint system $(A^\star, C^\star)$ is stabilizable.*

Proof. Note the equivalence of the following statements:

$$\begin{aligned} &(A, C) \text{ is detectable,} \\ \iff &\forall s \in \overline{\mathbb{C}^+}, \text{rank}\left(\begin{bmatrix} sI-A \\ C \end{bmatrix}\right) = n, \\ \iff &\forall s \in \overline{\mathbb{C}^+}, \text{rank}\left(\begin{bmatrix} s^\star I - A^\star & C^\star \end{bmatrix}\right) = n, \\ \iff &\forall s \in \overline{\mathbb{C}^+}, \text{rank}\left(\begin{bmatrix} sI - A^\star & C^\star \end{bmatrix}\right) = n, \end{aligned}$$

where complex conjugation is a surjection when defined on the left half complex plane. ■

Theorem 5.39. *Consider the LTI system (A, b, c^T), i.e.:*

$$\begin{aligned} \dot{x} &= Ax + bu, \\ y &= c^T x \end{aligned}$$

where A, b, c^T are given by:

$$A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-2} & -a_{n-1} \end{bmatrix}, \qquad b = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix},$$
$$c^T = \begin{bmatrix} c_0 & c_1 & c_2 & \cdots & c_{n-2} & c_{n-1} \end{bmatrix}$$

Then the transfer function of (A, b, c^T) is:

$$\begin{aligned} H(s) &= c^T(sI - A)^{-1}b \\ &= \frac{c_{n-1}s^{n-1} + \cdots + c_1 s + c_0}{s^n + a_{n-1}s^{n-1} + \cdots + a_1 s + a_0} \end{aligned}$$

Proof. Observe that $(sI - A)^{-1}b$ is simply the n-th column of $(sI - A)^{-1}$. Let its entries be denoted by $x_1, \cdots, x_n$. Then, we have:

$$\underbrace{\begin{bmatrix} s & -1 & 0 & \cdots & 0 & 0 \\ 0 & s & -1 & \cdots & 0 & 0 \\ 0 & 0 & s & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & s & -1 \\ a_0 & a_1 & a_2 & \cdots & a_{n-2} & s + a_{n-1} \end{bmatrix}}_{= sI - A} \cdot \underbrace{\begin{bmatrix} \cdots & \cdots & \cdots & \cdots & \cdots & x_1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & x_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots & x_3 \\ & & & & & \\ \cdots & \cdots & \cdots & \cdots & \cdots & \vdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & x_{n-1} \\ \cdots & \cdots & \cdots & \cdots & \cdots & x_n \end{bmatrix}}_{= (sI - A)^{-1}} = I$$

$$\Rightarrow \begin{bmatrix} s & -1 & 0 & \cdots & 0 & 0 \\ 0 & s & -1 & \cdots & 0 & 0 \\ 0 & 0 & s & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & s & -1 \\ a_0 & a_1 & a_2 & \cdots & a_{n-2} & s + a_{n-1} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}$$

Thus, we have the following system of linear equations:

$$\begin{cases} sx_1 - x_2 = 0, \\ sx_2 - x_3 = 0, \\ \quad \vdots \\ sx_{n-1} - x_n = 0, \\ a_0x_1 + a_1x_2 + \cdots a_{n-2}x_{n-1} + (s + a_{n-1})x_n = 1 \end{cases},$$

$$\Rightarrow \begin{cases} x_2 = sx_1, \\ x_3 = sx_2 = s^2x_1, \\ \quad \vdots \\ x_n = sx_{n-1} = \cdots s^{n-1}x_1, \\ a_0x_1 + a_1x_2 + \cdots a_{n-2}x_{n-1} + (s + a_{n-1})x_n \\ \qquad = (a_0 + a_1s + \cdots + a_{n-1}s^{n-1} + s^n)x_1 = 1 \end{cases},$$

$$\Rightarrow x_i = \frac{s^{i-1}}{s^n + a_{n-1}s^{n-1} + \cdots + a_1s + a_0}, \qquad i = 1, \cdots, n$$

Substituting back into our expression for $H(s)$, we find that:

$$\begin{aligned} H(s) &= c^T(sI - A)^{-1}b \\ &= \begin{bmatrix} c_0 & c_1 & \cdots & c_{n-1} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \\ &= \begin{bmatrix} c_0 & c_1 & \cdots & c_{n-1} \end{bmatrix} \cdot \frac{1}{s^n + a_{n-1}s^{n-1} + \cdots + a_1s + a_0} \begin{bmatrix} 1 \\ s \\ \vdots \\ s^{n-1} \end{bmatrix} \\ &= \frac{c_{n-1}s^{n-1} + \cdots + c_1s + c_0}{s^n + a_{n-1}s^{n-1} + \cdots + a_1s + a_0} \end{aligned}$$

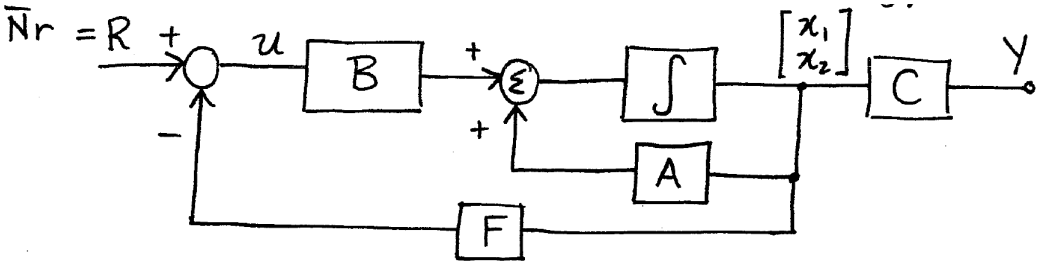
■

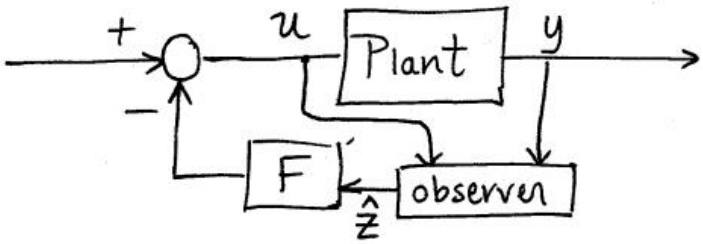
Remark. The transfer function does not necessarily reveal all the eigenvalues of A (and thus does not necessarily reveal all uncontrollable / unobservable modes), since pole-zero cancellation may have occurred.

5.9 Lecture 22

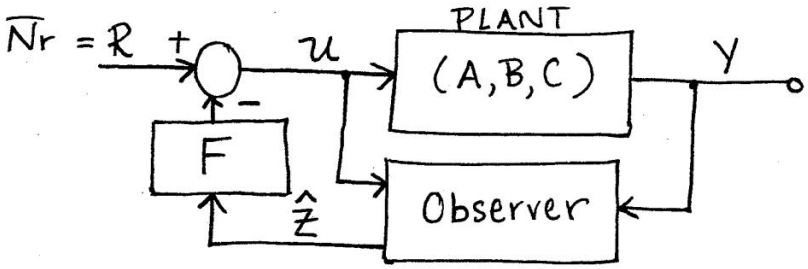
Observer Design:

In previous sections, we discussed how controllable (or stabilizable) LTI systems can be steered from one point in space and time to another via state feedback (e.g. $u = Fx$), as shown below:





Often, the state itself is inaccessible for direct measurement, and must be somehow estimated. In this section, we discuss how a signal reconstruction device, called an **observer**, can be designed to estimate these inaccessible states. Moreover, we show that *if the LTI system is observable, the observer can be designed such that its state estimation error asymptotically approaches 0 as* $t \to \infty$.



Consider an LTI plant $R : (A, B, C)$, with dynamics as shown below:

$$\dot{x} = Ax + Bu,$$
$$y = Cx,$$

where, as before, u, x, y denote the input, state, and output, respectively. We can then construct a full-order observer (i.e. an observer with states of the same dimensions as those in the original system). The observer takes as its input the inputs and outputs of the original system, and attempts to evolve its states in such a way that the difference ("error") between the observer state and system state vanishes asymptotically. Mathematically, we wish to choose the observer's dynamics:

$$\dot{\hat{z}} = M\hat{z} + Nu + Ty,$$

where $M \in \mathbb{R}^{n\times n}$, $N \in \mathbb{R}^{n\times n_i}$, $T \in \mathbb{R}^{n\times n_o}$, such that the *error*:

$$e \equiv \hat{z} - x$$

approaches 0 asymptotically as $t \to \infty$.

Observer with State Feedback—The Separation Theorem:

To see what choice of M, N, T would allow this to occur, we consider the system and observer dynamics together, as follows:

$$\begin{bmatrix}\dot{x}\\ \dot{\hat{z}}\end{bmatrix} = \begin{bmatrix}A & O\\ TC & M\end{bmatrix}\begin{bmatrix}x\\ \hat{z}\end{bmatrix} + \begin{bmatrix}B\\ N\end{bmatrix}u$$

Now, suppose we apply a *negative feedback based on the observer estimate*, i.e. $u = -F\hat{z} + r$, where v is known as the *reference* or *auxiliary input*. Applying this input, and using the change of variables $e \equiv \hat{z} - x$, we have:

$$\begin{aligned}
&\because \begin{bmatrix}\dot{x}\\ \dot{\hat{z}}\end{bmatrix} = \begin{bmatrix}A & -BF\\ TC & M-NF\end{bmatrix}\begin{bmatrix}x\\ \hat{z}\end{bmatrix} + \begin{bmatrix}B\\ N\end{bmatrix}v, \qquad \text{and } \begin{bmatrix}x\\ \hat{z}\end{bmatrix} = \begin{bmatrix}I & 0\\ I & I\end{bmatrix}\begin{bmatrix}x\\ e\end{bmatrix},\\
&\Rightarrow \begin{bmatrix}\dot{x}\\ \dot{e}\end{bmatrix} = \begin{bmatrix}I & 0\\ I & I\end{bmatrix}^{-1}\begin{bmatrix}A & -BF\\ TC & M-NF\end{bmatrix}\begin{bmatrix}I & 0\\ I & I\end{bmatrix}\begin{bmatrix}x\\ e\end{bmatrix} + \begin{bmatrix}I & 0\\ I & I\end{bmatrix}^{-1}\begin{bmatrix}B\\ N\end{bmatrix}v\\
&\qquad = \begin{bmatrix}I & 0\\ -I & I\end{bmatrix}\begin{bmatrix}A-BF & -BF\\ TC+M-NF & M-NF\end{bmatrix}\begin{bmatrix}x\\ e\end{bmatrix} + \begin{bmatrix}B\\ -B+N\end{bmatrix}v\\
&\qquad = \begin{bmatrix}A-BF & -BF\\ -(A-TC)+M+(B-N)F & M+(B-N)F\end{bmatrix}\begin{bmatrix}x\\ e\end{bmatrix} + \begin{bmatrix}B\\ -B+N\end{bmatrix}v
\end{aligned}$$

Since *we want the error to decay exponentially regardless of the state or input*, set:

$$\begin{aligned}
-(A-TC) + M + (B-N)F &= O,\\
-B + N &= O,\\
\sigma\big(M + (B-N)F\big) &\subset \mathbb{C}^-
\end{aligned}$$

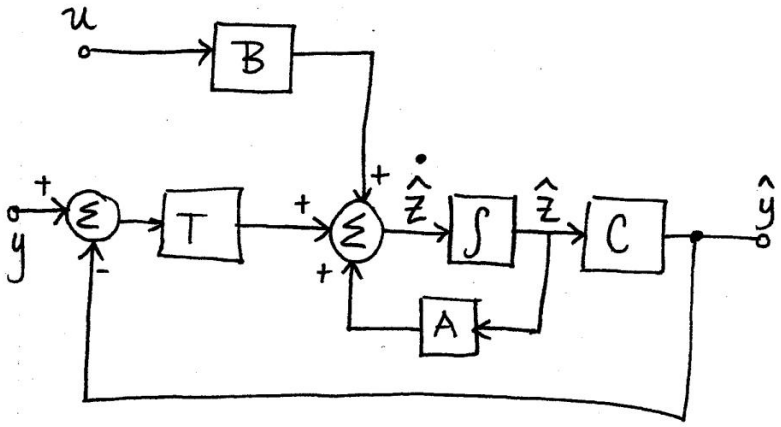
Rearranging terms, we have:

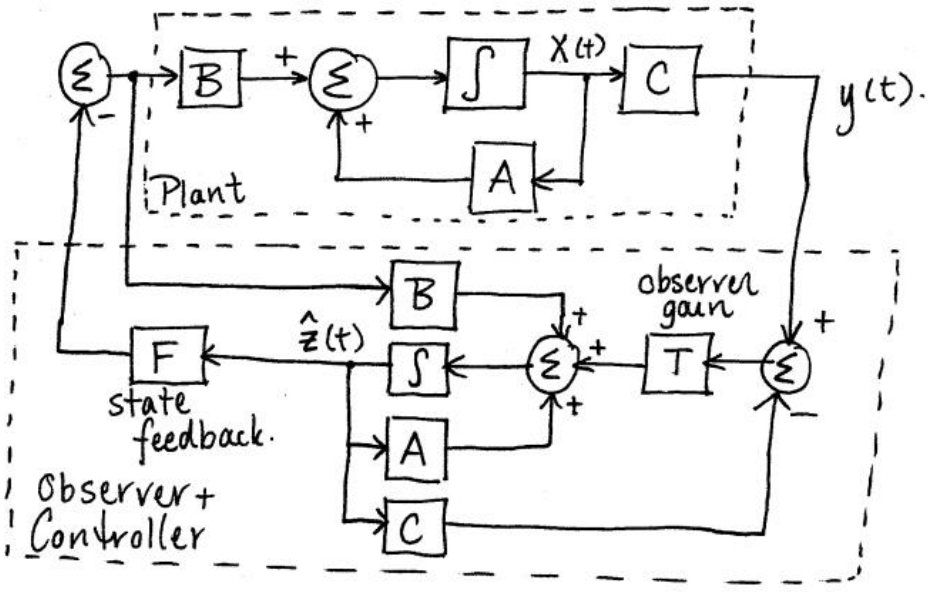
$$\begin{aligned}
M &= A - TC,\\
N &= B,\\
\sigma(A-TC) &\subset \mathbb{C}^-
\end{aligned}$$

Notice that *the final requirement can be achieved if* (A, C) *is observable*. In that case, by substituting back the above choices of M, N, T, we find that our desired full-order observer must evolve according to the dynamics:

$$\begin{aligned}\dot{\hat{z}} &= (A - TC)\hat{z} + Bu + TCx \\ &= A\hat{z} + Bu + T(y - C\hat{z}), \\ \Rightarrow \begin{bmatrix}\dot{x} \\ e\end{bmatrix} &= \begin{bmatrix}A - BF & -BF \\ O & A - TC\end{bmatrix}\begin{bmatrix}x \\ e\end{bmatrix} + \begin{bmatrix}B \\ O\end{bmatrix} v,\end{aligned}$$

which can be physically implemented by feeding the difference between the outputs of the system and observer back into the observer as an input; a figure is provided below.





The above derivation shows that the set of eigenvalues of the composite system is simply:

$$\sigma(A - BF) \cup \sigma(A - TC).$$

Thus, *the problem of arbitrarily assigning the closed-loop poles of a system using feedback can be "separated" into two parts:*

1. Designing an observer to provide a set of asymptotically-accurate state estimates, by designing the *observer (estimation) poles* $\sigma(A - TC)$, and

2. Designing a pole-assigning state feedback matrix as though the true states were available for direct measurement, by designing the*closed loop poles* $\sigma(A - BF)$.

Notice that $\sigma(A-TC) = \sigma((A-TC)^T) = \sigma(A^T - C^T T^T)$ whenever A is a square matrix, and C, T are of appropriate dimension. Thus, the pole placement algorithm for observer design is identical to the pole placement algorithm for state feedback design, with A, B, F replaced by A^T, C^T, T^T, respectively. For this reason, the standard observer configuration is said to be the *dual* of the state feedback configuration.

Remark (**Guidelines for Pole Placement**). The following guidelines are useful for deciding where the closed-loop poles, $\sigma(A-BF)$ and the closed-loop (estimator poles) $\sigma(A-TC)$ should be placed:

1. Closed-loop (Estimator) Poles:

   - The larger the gain, the larger the control input—this is because $u = -Fx$, i.e. the control input u is proportional to the gain matrix F.
   - The more the poles are moved from open-loop systems (with "A") to closed-loop systems (with "$A - BF$"), the larger the gain matrix F. This arises from the fact that the following is a sequence of continuous mappings:

   $$\begin{aligned} &\text{Elements in square matrix } A \\ \longrightarrow\ &\text{The characteristic polynomial of } A, \text{i.e. } \chi_A(s) \\ \longrightarrow\ &\text{The roots of } \chi_A(s), \text{i.e. the eigenvalues of } A \end{aligned}$$

2. Estimator Poles:

   - Estimator poles are generally chosen to be to the left (on the complex plane, i.e. with a more negative real part) than the controller poles. This is because we wish the estimator error to have a faster rate of decay compared with the desired dynamics.
   - In practice, however, it is often a bad idea to move estimator poles too far to the loeft, since this increases the bandwidth of the estimator, causing more sensor noise to pass on to the control actuator.

Example (**Inverted Pendulum with Disturbance**). Consider a slightly modified version of the inverted pendulum example given in Lecture 10, where θ is the angle of deviation of the pendulum from its vertical (pointing upwards) position, α denotes some constant disturbance, and the states are defined as:

$$(x_1, x_2, x_3) = (\theta, \dot{\theta}, d),$$

with dynamics given by:

$$\begin{aligned}
\dot{x}_1 &= x_2, \\
\dot{x}_2 &= \Omega^2 x_1 - \alpha x_2 + x_3 + u, \\
\dot{x}_3 &= 0. \\
y &= x_1
\end{aligned}$$

In matrix form, we have:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & 1 & 0 \\ \Omega^2 & -\alpha & 1 \\ 0 & 0 & 0 \end{bmatrix}}_{=A} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \underbrace{\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}}_{=B} u,$$

$$y = \underbrace{\begin{bmatrix} 1 & 0 & 0 \end{bmatrix}}_{=C} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

Thus, if we design the observer dynamics with $T = \begin{bmatrix} T_1 & T_2 & T_3 \end{bmatrix}^T$, we have:

$$\begin{aligned}
\dot{\hat{z}} &= A\hat{z} + Bu + T(y - C\hat{z}) = (A - TC)\hat{z} + Bu + Ty \\
&= \begin{bmatrix} -T_1 & 1 & 0 \\ \Omega^2 - T_2 & -\alpha & 1 \\ -T_3 & 0 & 0 \end{bmatrix} \hat{z} + \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} u + \begin{bmatrix} T_1 \\ T_2 \\ T_3 \end{bmatrix} y
\end{aligned}$$

Example. Consider the following system:

$$\begin{aligned}
\dot{x} &= \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & -1 \end{bmatrix} x + \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} u, \\
y &= \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} x
\end{aligned}$$

1. Design an observer with poles at -4 and $-4 \pm j2$.

2. Design a state feedback so that the closed loop poles are located at -2 and $-2 \pm j2$.

Solution:

1. The poles of the observer the eigenvalues of $A-TC$, which, if $T = \begin{bmatrix} T_1 & T_2 & T_3 \end{bmatrix}^T \in \mathbb{R}^{1\times 3}$, equals:

$$\begin{aligned} A - TC &= \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & -1 \end{bmatrix} - \begin{bmatrix} T_1 \\ T_2 \\ T_3 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 0 & 0 & -T_1 \\ 1 & 0 & -T_2 \\ 0 & 1 & -1-T_3 \end{bmatrix} \end{aligned}$$

with characteristic polynomial:

$$\begin{aligned} \chi_{A-TC}(s) &= \det\left(\begin{bmatrix} s & 0 & T_1 \\ -1 & s & T_2 \\ 0 & -1 & s+1+T_3 \end{bmatrix} \right) \\ &= s\big(s(s+1+T_3)+T_2\big)+T_1 \\ &= s^3 + (1+T_3)s^2 + T_2 s + T_1 \end{aligned}$$

Now, to place the poles at -4 and -4 ± 2, we want this characteristic polynomial to be:

$$\begin{aligned} &(s+4)(s-(-4+j2))(s-(-4-j2)) \\ =&(s+4)((s+4)^2+4) = (s+4)(s^2+8s+20) \\ =&s^3+12s^2+52s+80 \end{aligned}$$

Comparing the two polynomials, we see that we should set:

$$T = \begin{bmatrix} T_1 \\ T_2 \\ T_3 \end{bmatrix} = \begin{bmatrix} 80 \\ 52 \\ 11 \end{bmatrix}$$

2. The closed-loop poles the eigenvalues of $A - BF$, which, if $F = \begin{bmatrix} F_1 & F_2 & F_3 \end{bmatrix} \in \mathbb{R}^3$, equals:

$$\begin{aligned} A - BF &= \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & -1 \end{bmatrix} - \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} F_1 & F_2 & F_3 \end{bmatrix} \\ &= \begin{bmatrix} -F_1 & -F_2 & -F_3 \\ 1 & 0 & 0 \\ 0 & 1 & -1 \end{bmatrix} \end{aligned}$$

with characteristic polynomial:

$$\begin{aligned} \chi_{A-BF}(s) &= \det\left(\begin{bmatrix} s+F_1 & F_2 & F_3 \\ -1 & s & 0 \\ 0 & -1 & s+1 \end{bmatrix} \right) \\ &= (s+F_1)s(s+1) + F_2(s+1) + F_3 \\ &= s^3 + (F_1+1)s^2 + (F_1+F_2)s + (F_2+F_3) \end{aligned}$$

Now, to place the poles at $-$ and -4 ± 2, we want this characteristic polynomial to be:

$$\begin{aligned}&(s+2)(s-(-2+j2))(s-(-2-j2))\\ =&(s+2)((s+2)^2+4) = (s+2)(s^2+4s+8)\\ =&s^3+6s^2+16s+16\end{aligned}$$

Comparing the two polynomials, we see that we should set:

$$F = \begin{bmatrix}F_1 & F_2 & F_3\end{bmatrix} = \begin{bmatrix}5 & 11 & 5\end{bmatrix}$$

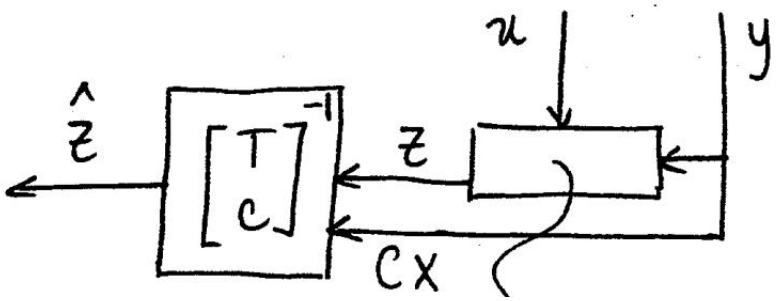
Reduced Order Observer Design:

In the above analysis, we constructed *full-order observers* for LTI systems, i.e. observers with states of the same dimension as the states in the original LTI system. However, this may be needlessly complicated, especially if several of the states are directly measured in y.

To be more specific, since $y = Cx$, where $C \in \mathbb{R}^{p\times n}$ we can derive $\text{rank}(C)$ linearly independent relations among the different dimensions of the state x at any given time. We can assume C has full row rank, i.e. $\text{rank}(C) = p$, since if not, we can simply remove redundant outputs from our system representation via Gaussian elimination. Now, instead of designing a full-order observer with n states, we can instead design a *reduced-order observer* with $n - \text{rank}(C) = n - p$ states, which, together with the direct observations $y = Cx$, allow us to reconstruct an estimate of the system state with asymptotically decaying error. The general idea is to design an observer some $T = \mathbb{R}^{(n-p)\times n}$ such that:

$$\begin{bmatrix}T\\C\end{bmatrix} \in \mathbb{R}^{n\times n}$$

is invertible. Note that the matrix $T \in \mathbb{R}^{(n-p)\times p}$ here maps the output of the partial state observer to the estimate of the state, as shown in the figure below. It is *not* the output feedback used in the full observer case; for the partial state observe, we use "L" instead, as described below.



As with the full observer case, we consider the estimator dynamics to be of the form:

$$\dot{\hat{z}} = Mz + Nu + Ly$$

for some $M \in \mathbb{R}^{p\times p}, N \in \mathbb{R}^{p\times n_i}, L \in \mathbb{R}^{p\times n_o}$. (Note that, since the estimator state z and system state x now have different dimensions, M is no longer $n \times n$.) We wish to choose M, N, L such that the *error*:

$$e \equiv \hat{z} - Tx$$

asymptotically approaches 0 as $t \to \infty$. To do so, we retrace the steps taken in the derivation of the full-state observer. Since:

$$\begin{aligned}
\dot{x} &= Ax + Bu, \\
y &= Cx, \\
\dot{\hat{z}} &= M\hat{z} + Nu + Ly,
\end{aligned}$$

and $e = \hat{z} - Tx$, we have:

$$\begin{aligned}
\because \begin{bmatrix} \dot{x} \\ \dot{\hat{z}} \end{bmatrix} &= \begin{bmatrix} A & O \\ LC & M \end{bmatrix} \begin{bmatrix} x \\ \hat{z} \end{bmatrix} + \begin{bmatrix} B \\ N \end{bmatrix} u, \qquad \text{and } \begin{bmatrix} x \\ \hat{z} \end{bmatrix} = \begin{bmatrix} I & O \\ T & I \end{bmatrix} \begin{bmatrix} x \\ e \end{bmatrix}, \\
\Rightarrow \begin{bmatrix} \dot{x} \\ \dot{e} \end{bmatrix} &= \begin{bmatrix} I & 0 \\ T & I \end{bmatrix}^{-1} \begin{bmatrix} A & O \\ LC & M \end{bmatrix} \begin{bmatrix} I & 0 \\ T & I \end{bmatrix} \begin{bmatrix} x \\ e \end{bmatrix} + \begin{bmatrix} I & 0 \\ T & I \end{bmatrix}^{-1} \begin{bmatrix} B \\ N \end{bmatrix} u \\
&= \begin{bmatrix} I & 0 \\ -T & I \end{bmatrix} \begin{bmatrix} A & O \\ LC + MT & M \end{bmatrix} \begin{bmatrix} x \\ e \end{bmatrix} + \begin{bmatrix} I & 0 \\ -T & I \end{bmatrix} \begin{bmatrix} B \\ N \end{bmatrix} u \\
&= \begin{bmatrix} A & O \\ MT - TA + LC & M \end{bmatrix} \begin{bmatrix} x \\ e \end{bmatrix} + \begin{bmatrix} B \\ -TB + N \end{bmatrix} u
\end{aligned}$$

In other words:

$$\begin{aligned}
\dot{x} &= Ax + Bu, \\
\dot{e} &= (MT - TA + LC)x + Me + (N - TB)u.
\end{aligned}$$

Since we wish to design M, N, L such that $e \to 0$ as $t \to \infty$, the above dynamics suggests choosing M, N, L such that $\dot{e} = Me$, with M to be asymptotically stable (i.e. with $\sigma(M) \in \mathbb{C}^-$). In other words, we want:

1. $N = TB$.
2. $MT - TA + LC = O$.
3. $\sigma(M) \in \mathbb{C}^-$

The following design steps provide one method of achieving the above objectives:

1. Choose M such that $\sigma(M) \in \mathbb{C}^-$.

2. Choose L.

3. Solve for T from $MT - TA + LC = O$.

4. Solve for N in $N = TB$.

5. Check that $\begin{bmatrix} T \\ C \end{bmatrix}$ is invertible, e.g. by showing that:

$$\det\left(\begin{bmatrix} T \\ C \end{bmatrix}\right) \neq 0.$$

While this design procedure may appear reasonable for simple systems, it is *not recommended* in general, since it gives the designer no control over the matrix $\begin{bmatrix} T \\ C \end{bmatrix}$. Thus, if $\begin{bmatrix} T \\ C \end{bmatrix}$ is close to being singular, its inverse would result in a huge gain in the backward loop.

Consider instead the following design.

Reduced Order Observer Design—Algorithm:

1. Find an equivalent representation for the plant as follows; as before, we assume C to be of full column rank (i.e. $\text{rank}(C) = p$):

$$\Sigma : \begin{cases} \dot{x} = Ax + Bu, \\ y = Cx \end{cases}, \qquad \begin{cases} \dot{\overline{x}} = \overline{A}\overline{x} + \overline{B}\overline{u}, \\ \overline{y} = \overline{C}\overline{x} \end{cases}$$

where $\overline{x} = Sx$ and:

$$\begin{aligned} \overline{A} &= S^{-1}AS \equiv \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \\ \overline{B} &= S^{-1}B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} \\ \overline{C} &= CS = \begin{bmatrix} C_1 & O \end{bmatrix} \end{aligned}$$

for some invertible $S \in \mathbb{R}^{n\times n}$, and invertible $C_1 \in \mathbb{R}^{p\times p}$. That $\overline{C} = CS = \begin{bmatrix} C_1 & O \end{bmatrix}$, with C_1 invertible, can always be achieved by choosing the invertible $S \in \mathbb{R}^{n\times n}$, when multiplied to the right of C, to rearrange the columns of C in such a way that the first p columns become linearly independent, and the remaining columns are annihilated. (In linear algebra, we call this the *column echelon form of* C). By construction, $A_{11} \in \mathbb{R}^{p\times p}, B_1 \in \mathbb{R}^{p\times m}, C_1 \in \mathbb{R}^{p\times p}$

2. Choose T, M:

Design an observer for $\overline{x}$, i.e. construct a state estimate $\hat{z}$ such that $\hat{z} \to x$ asymptotically as $t \to \infty$. Specifically, we choose $T = \begin{bmatrix} T_1 & I \end{bmatrix}$ for some $T_1 \in \mathbb{R}^{(n-p)\times p}$. This is done to force $\begin{bmatrix} T \\ C \end{bmatrix}$ to assume the form:

$$\begin{bmatrix} T \\ \overline{C} \end{bmatrix} = \begin{bmatrix} -T_1 & I \\ C_1 & O \end{bmatrix},$$

which would then always be invertible, as I and C_1 both have full row rank. It remains to solve for T_1, which we can do using the criterion $MT - T\overline{A} + L\overline{C} = O$:

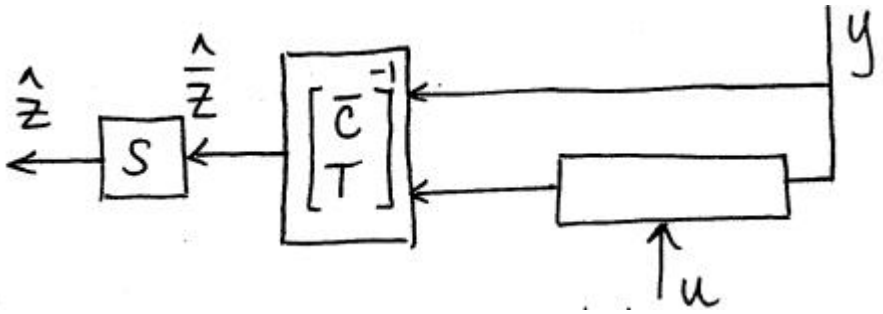
$$\begin{aligned} O &= MT - T\overline{A} + L\overline{C} \\ &= M\begin{bmatrix} -T_1 & I \end{bmatrix} - \begin{bmatrix} -T_1 & I \end{bmatrix}\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} + L\begin{bmatrix} C_1 & O \end{bmatrix} \\ &= \begin{bmatrix} -MT_1 + T_1A_{11} - A_{21} + LC_1 & M + T_1A_{12} - A_{22} \end{bmatrix} \end{aligned}$$

In particular, $M = A_{22} - T_1A_{12}$, and we wish for M to have eigenvalues strictly on the left half complex plane. Notice that *this is simply a standard pole placement problem*; that is, if (A_{22}, A_{12}) is observable, then there exists some T_1 such that $\sigma(M) = \sigma(A_{22} - T_1A_{12})$ can be placed wherever we want.

We conclude our analysis by showing that, *if* (A, C) *is observable*, then (A_{22}, A_{12}) is indeed observable. This is because, for each $s \in \mathbb{C}$:

$$\begin{aligned} n &= \operatorname{rank}\left(\begin{bmatrix} sI - A_{11} & -A_{12} \\ -A_{21} & sI - A_{22} \\ C_1 & O \end{bmatrix}\right) = \operatorname{rank}\left(\begin{bmatrix} O & -A_{12} \\ O & sI - A_{22} \\ I & O \end{bmatrix}\right) \\ &= p + \operatorname{rank}\left(\begin{bmatrix} -A_{12} \\ sI - A_{22} \end{bmatrix}\right) = p + \operatorname{rank}\left(\begin{bmatrix} sI - A_{22} \\ A_{12} \end{bmatrix}\right) \\ \Rightarrow \operatorname{rank}&\left(\begin{bmatrix} sI - A_{22} \\ A_{12} \end{bmatrix}\right) = n - p, \end{aligned}$$

as applying elementary row operations to a matrix does not affect its rank. Thus, $\begin{bmatrix} sI - A_{22} \\ A_{12} \end{bmatrix}$ has full column rank.



3. Solve for L:

   From the above derivation, we have:

$$\begin{aligned} &-MT_1 + T_1A_{11} - A_{21} + LC_1 = O, \\ \Rightarrow &L = (MT_1 - T_1A_{11} + A_{21})C_1^{-1} \end{aligned}$$

4. Solve for N:

   Finally, we have:

$$\begin{aligned} N = T\overline{B} &= \begin{bmatrix} -T_1 & O \end{bmatrix} \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} \\ &= -T_1B_1 + B_2 \end{aligned}$$

Example. Design a reduced order observer for the system:

$$\begin{aligned} \dot{x} &= \begin{bmatrix} 0 & 1 \\ -2 & 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u, \\ y &= \begin{bmatrix} 1 & 0 \end{bmatrix} x \end{aligned}$$

Solution :

Observe that $n = 2$, $p = 1$, and C is already in column echelon form.

1. Since C is already in column echelon form with $p = 1$, we have $C_1 = 1$.

2. Let $T = \begin{bmatrix} -T_1 & 1 \end{bmatrix}$, and observe that:

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -2 & 0 \end{bmatrix}$$

   Thus, if we choose a particular stable M, e.g. $M = -1$, we can solve for T_1 as follows:

$$\begin{aligned} 1 = M &= A_{22} - T_1A_{12} = 0 - t_1 \cdot 1 = -t_1, \\ \Rightarrow T_1 &= 1. \end{aligned}$$

3. Solve for L:

$$\begin{aligned} L &= (MT_1 - T_1A_{11} + A_{21})C^{-1} \\ &= (-1 \cdot 1 - 1 \cdot 0 + (-2)) \cdot 1^{-1} \\ &= -3. \end{aligned}$$

4. Solve for N:

$$N = -T_1B_1 + B_2 = -1 \cdot 0 + 1 = 1.$$

Chapter 6

Additional Topics

6.1 Lecture 11

In this final lecture, we combine everything we have learned in previous lectures and examine a classical problem in control theory—the linear quadratic regulator (LQR). This lecture draws largely from Professor Daniel Liberzon's text "Calculus of Variations and Optimal Control, A Concise Introduction" [6].

Note. We will assume $t_0 = 0$, and positive state feedback, throughout the following derivations.

Finite-Horizon LQR Problem—Riccati Differential Equation Method

Definition 6.1 (**Finite-Horizon LQR Problem**). *The* ***finite-horizon linear quadratic optimal control problem*** *is defined as follows—Given the system:*

$$\begin{aligned} \dot{x} &= Ax + Bu, \qquad x(0) = x_0, \\ y &= Cx, \end{aligned}$$

evolving in the time interval $[0, t_1]$ *with* $x \in \mathbb{R}^n$, $u \in \mathbb{R}^{n_i}$, *and* (A, B) *controllable,* (A, C) *observable, find the input function* $u(\cdot) : [0, t_1]$ *that minimizes a* ***quadratic cost functional****:*

$$J\big(u(\cdot)\big) = \int_0^{t_1} \big[x^\star Q x + u^\star R u\big] dt + x(t_1)^\star S x(t_1)$$

satisfying $Q \equiv C^\star C \geq 0$, $R > 0$ *for each* $t \in [0, t_1]$, *and* $S \geq 0$. *(The argument* t *is abbreviated in the integrand, for ease of notation).*

Just as the Lyapunov equation can be motivated by a quadratic value function, here we attempt to rewrite the final cost $x(t_1)^T S x(t_1)$ in terms of some initial cost, plus an integrated term. In particular, consider the following theorem.

Theorem 6.2 (**Finite-Horizon LQR Solution: RDE Method**). *Let $P(t) \in \mathbb{R}^{n\times n}$ be the solution to the final state Riccati differential equation (RDE):*

$$\dot{P} + A^\star P + PA - PBR^{-1}B^\star P + Q = 0, \qquad P(t_1) = S.$$

Then the optimal control to the finite-horizon LQR problem is the linear time-varying control:

$$u_{\text{opt}}(t) = -R^{-1}B^\star P x(t)$$

where the argument t is hidden in R, B, P, for convenience of notation. The corresponding optimal cost is:

$$J(u_{\text{opt},[0,t]}) = x_0^\star P(0) x_0$$

Proof. By directly expanding the cost function, we have:

$$\begin{aligned}
J(u_{[0,t_1]}) &= \int_0^{t_1} (x^\star Q x + u^\star R u)\, dt + x(t_1)^\star S x(t_1) \\
&= \int_0^{t_1} (x^\star Q x + u^\star R u)\, dt + x(0)^\star P(0) x(0) + \int_0^{t_1} \frac{d}{dt}(x^\star P x) dt \\
&= \int_0^{t_1} (x^\star Q x + u^\star R u)\, dt + x(0)^\star P(0) x(0) \\
&\qquad + \int_0^{t_1} \big[(Ax+Bu)^\star P x + x^\star P(Ax+Bu) + x^\star \dot{P} x\big] dt \\
&= \int_0^{t_1} (x^\star Q x + u^\star R u)\, dt + x(0)^\star P(0) x(0) \\
&\qquad + \int_0^{t_1} \big[x^\star A^\star P x + x^\star P A x + u^\star B^\star P x + x^\star P B u + x^\star \dot{P} x\big] dt \\
&= x(0)^\star P(0) x(0) + \int_0^{t_1} x^\star \big(\dot{P} + A^\star P + PA + Q - PBR^{-1}B^\star P\big) x\, dt \\
&\qquad + \int_0^{t_1} \big[(u + R^{-1}B^\star P x)^\star R (u + R^{-1}B^\star P x)\big] dt \\
&= x(0)^\star P(0) x(0) + \int_0^{t_1} \big[(u + R^{-1}B^\star P x)^\star R (u + R^{-1}B^\star P x)\big] dt
\end{aligned}$$

where, in the second-to-last step, we have completed the square inside the second integral, to force the integrand into a quadratic form.

The optimal control and cost are thus:

$$\begin{aligned}
u_{opt} &\equiv \arg\min_{u_{[t_0,t_1]}} J(u_{[t_0,t_1]}) = -R^{-1}B^\star P x, \\
\Rightarrow\ J(u_{opt}) &= \min_{u_{[t_0,t_1]}} J(u_{[t_0,t_1]}) = x_0^\star P x_0
\end{aligned}$$

Since $J(u_{[t_0,t_1]}) \geq 0$ for any arbitrary x_0, we have $P_0 \geq 0$. ■

Remark. In general, Riccati differential equations are mathematically defined as the first-order differential equations that are quadratic, i.e. of the form:

$$\dot{x}(t) = p_0(t) + p_1(t)x(t) + p_2(t)x^2(t),$$

where $p_2(t)$ is a non-zero function. However, here, for the purposes of solving the LQR problem, we will restrict ourselves the particular matrix form of the RDE shown above.

Infinite-Horizon LQR Problem

Below, we state the infinite-horizon analogue of the finite-horizon LQR problem.

Definition 6.3 (**Infinite-Horizon LQR Problem**). *The* ***infinite-horizon linear quadratic optimal control problem*** *is defined as follows—Given the system:*

$$\begin{aligned} \dot{x} &= Ax + Bu, \qquad x(0) = x_0, \\ y &= Cx, \end{aligned}$$

evolving in the time interval $[0, \infty)$ *with* $x \in \mathbb{R}^n$, $u \in \mathbb{R}^{n_i}$, *and* (A, B) *controllable,* (A, C) *observable, find the input function* $u_{[0,\infty)}$ *that minimizes the quadratic cost:*

$$J\big(u(\cdot)\big) = \int_0^\infty \big[x^\star Q x + u^\star R u\big] dt$$

satisfying $Q \equiv C^\star C \geq 0$, $R > 0$ *for each* $t \in [0, \infty)$. *(Again, the argument* t *is abbreviated in the integrand, for ease of notation).*

Remark. The infinite-horizon LQR problem is simply its finite-horizon LQR counterpart with $t_1 \to \infty$ and $S = 0$. The reason we set $S = 0$ is because, in general, for the infinite-horizon quadratic cost to be finite, we require the state and input to converge to 0.

Our strategy to solving the infinite-horizon LQR problem will be to observe its similarities with its finite-horizon LQR problem. In particular, since we wish to take $t_1 \to \infty$ in the infinite-horizon variant of the problem, let us regard t_1 as a parameter to be varied, rather than fixed (as the was case for the finite-horizon LQR). In particular, we will denote by $P(t, t_1)$ the solution at time t of the RDE:

$$\begin{aligned} &\dot{P} + A^\star P + PA - PBR^{-1}B^\star P + Q = 0, \\ &P(t_1) = 0. \end{aligned}$$

and let the corresponding finite-horizon LQR optimal control and minimum cost be:

$$\begin{aligned} u^{t_1}_{opt}(t) &= -R^{-1}B^\star P(t, t_1)x(t), \qquad \forall t \in [0, t_1], \\ J^{t_1}(u^{t_1}_{opt,[0,t_1]}) &= x_0^\star P(t, t_1)x_0. \end{aligned}$$

It is natural to conjecture that the infinite-horizon optimal cost and optimal control can be obtained by simply applying the limit $t_1 \to \infty$ to the above finite-horizon solutions:

$$u^{\infty}_{opt}(t) = -R^{-1}B^{\star}\left(\lim_{t_1\to\infty} P(t,t_1)\right)x(t), \qquad \forall t \in [0,\infty)$$
$$J^{\infty}(u^{\infty}_{opt,[0,\infty)}) = x_0^{\star}\left(\lim_{t_1\to\infty} P(t,t_1)\right)x_0.$$

The next theorem indicates that this conjecture is correct. To do so, we must establish several facts, such as the existence and positive-definiteness of the limit $\lim_{t_1\to\infty} P(t,t_1)$.

Theorem 6.4 (**Infinite-Horizon LQR Solution**). *Suppose we are given the LTI system described in the infinite-horizon LQR problem.*

1. *Consider the Riccati Differential Equation and the Algebraic Riccati Equation given below:*

$$\dot{P} = -PA - A^{\star}P - Q + PBR^{-1}B^{\star}P, \qquad P(t_1) = 0, \tag{6.1}$$
$$PA + A^{\star}P + Q - PBR^{-1}B^{\star}P = O. \tag{6.2}$$

   Let $P(0,t_1)$ be the solution of (6.1) at time 0. Then:

$$P \equiv \lim_{t_1\to\infty} P(0,t_1)$$

   exists and is a positive definite solution of (6.2).

2. *The optimal cost is:*

$$J(u_{\text{opt}}) = x_0^T P x_0,$$

   which equals the $t_1 \to \infty$ limit of the finite-horizon optimal cost $x_0^T P(0,t_1)x_0$. An optimal control that achieves this optimal cost is the linear time-invariant state feedback:

$$u_{\text{opt}}(t) = -R^{-1}B^{\star}Px(t),$$

   which is the limit as $t_1 \to \infty$ of the finite-horizon optimal feedback $\overline{u_{t_1}}(t) = -R^{-1}B^{\star}P(t,t_1)x(t)$ derived in the above section on the finite-horizon LQR problem.

3. *The corresponding closed-loop system $\dot{x_{\text{opt}}} = (A - BR^{-1}B^{\star}P)x_{\text{opt}}$ is exponentially stable.*

4. *P is the unique positive definite solution to (6.2).*

Proof.

1. We divide the proof for this portion of the theorem into four parts. First, we observe the behavior of $P(0,t_1)$ as $t_1 \to \infty$. Then, we show that $\lim_{t\to\infty} P(0,t_1)$ exists, and satisfies (6.2).

- Monotonicity of $x_0^\star P(0,t_1)x_0$ with respect to t_1:

  The finite-horizon optimal cost $x_0^\star P(0,t_1)x_0$ is a monotonically non-decreasing function of the final time t_1, since, if $t_2 > t_1$:

$$\begin{aligned} x_0^\star P(0,t_2)x_0 &= \int_0^{t_2} \left[x^\star Q x + u^\star R u\right] dt \geq \int_0^{t_1} \left[x^\star Q x + u^\star R u\right] dt \\ &= x_0^\star P(0,t_1)x_0 \end{aligned}$$

  Note that the trajectory $x(t)$ in the above two integrands is the same non-negative value, since we apply the same control in the interval $[t_0, t_1]$, and $Q \geq 0, R > 0$.

- Boundedness of $x_0^\star P(t_0,t_1)x_0$, given (A,B) controllable:

  Since (A,B) is controllable, there exists some time T and a control $u'_{[0,T]}$ that steers the LTI system from $(x_0, 0)$ to $(0,T)$. Let $u'_{[0,\infty)}$ denote the control that equals$u'_{[0,T]}$ on the interval $[0,T]$, and 0 afterwards. If we apply $u'_{[0,\infty)}$, we obtain a state trajectory that is identically zero for all $t \geq T$. Now, since $x_0^\star P(0,t_1)x_0$ is the optimal finite-horizon cost for any $t_1 \geq 0$, we have for each $t_1 \geq T$:

$$x_0^\star P(0,t_1)x_0 \leq \int_0^\infty \left[x^\star Q x + u'^\star R u'\right] dt = \int_0^T \left[x^\star Q x + u'^\star R u'\right] dt < \infty$$

  This establishes the boundedness of $x_0^\star P(0,t_1)x_0$ with respect to t_1.

- Existence and Positive Definiteness of $P \equiv \lim_{t_1\to\infty} P(0,t_1)$:

  The two claims above, combined with the Monotonic Sequence Convergence Theorem in analysis (see [8], Theorem 3.14), indicate that $\lim\limits_{t_1\to\infty} x_0^\star P(0,t_1)x_0$ exist for any choice of x_0. Here, we wish to show that, in fact, the matrix $\lim\limits_{t_1\to\infty} P(0,t_1)$ is also well-defined. To see this, let $e_i, e_j \in \mathbb{R}^n$ be any two distinct standard vectors in $\mathbb{R}^n$. Then:

$$\begin{aligned} e_i^\star P(0,t_1)e_i &= P_{ii}(0,t_1), \\ e_j^\star P(0,t_1)e_j &= P_{jj}(0,t_1), \\ (e_i+e_j)^\star P(0,t_1)(e_i+e_j) &= P_{ii}(0,t_1) + 2P_{ij}(0,t_1) + P_{jj}(0,t_1). \end{aligned}$$

  where $P_{kl}[0,t_1]$ denotes the (k,l)-th element of $P(0,t_1)$, for any $k,l \in \{1,\cdots,n\}$. Since the left-hand side of each of the above equalities converges as $t_1 \to \infty$, so does the right-hand side. Thus, $\lim_{t_1\to\infty} P_{ij}(0,t_1)$ exists for any $i,j \in \{1,\cdots,n\}$, and it automatically follows that so does:

$$P \equiv \lim_{t_1\to\infty} P(0,t_1).$$

One can think of the above limit as the matrix obtained by starting from the zero matrix at some very negative initial time $t_0 \to -\infty$, then propagating forward to time 0 along the RDE (6.1).

Observe that P is positive semi-definite, since $P(0,t_1)$ is positive semi-definite for each $t_1 \geq 0$, and $P(t,t_1)$ is continuous in t_1 (see Fundamental Theorem of Differential Equations, i.e. Theorem 3.4).

- $P \equiv \lim_{t_1\to\infty} P(0,t_1)$ satisfies (6.2):

  Applying $t_1 \to \infty$ to both sides of the RDE (6.1), we find that $\lim_{t_1\to\infty} \dot{P}(0,t_1)$ must also exist, and equal the zero matrix. Thus, P satisfies the ARE (6.2).

2. Observe that, for any control $u_{[0,\infty)}$ over the infinite time horizon, we have:

$$\begin{aligned} J(u_{[0,\infty)}) &= \int_0^\infty \left[x^\star Q x + u^\star R u\right] dt \geq \int_0^{t_1} \left[x^\star Q x + u^\star R u\right] dt \\ &\geq x_0^\star P(0,t_1) x_0. \end{aligned}$$

for any $t_1 \geq 0$. Taking $t_1 \to \infty$ on both sides, we have:

$$J(u_{[0,\infty)}) \geq x_0^\star P x_0.$$

Thus, the infinite-horizon cost must be at least $x_0^\star P x_0$.

Meanwhile, if we take the input to be $u_{opt}(t) \equiv -R^{-1}B^\star P x(t)$, we have the trajectory:

$$\dot{x} = Ax + Bu = (A - BR^{-1}B^\star P)x$$

Thus, over any finite horizon $[0,t)$:

$$\begin{aligned} J^{t_1}(u_{opt}) &= \int_0^{t_1} \left[x^\star Q x + u_{opt}^\star R u_{opt}\right] dt = \int_0^{t_1} x^\star (Q + PBR^{-1}B^\star P) x \, dt \\ &\leq \int_0^{t_1} x^\star (PA + A^\star P - 2PBR^{-1}B^\star P) x \, dt \\ &\leq -\int_0^{t_1} \left[x^\star (A - BR^{-1}B^\star P)^\star P x + x^\star P (A - BR^{-1}B^\star P) x\right] dt \\ &\leq -\int_0^{t_1} \left[\dot{x}^\star P x + x^\star P \dot{x}\right] dt \\ &= -\int_0^{t_1} \frac{d}{dt}(x^\star P x)\, dt \\ &\leq x_0^\star P x_0 - x(t_1)^\star P x(t_1) \\ &\leq x_0^\star P x_0 \end{aligned}$$

Taking $t_1 \to \infty$ on both sides, we have:

$$J(u_{opt}) \leq x_0^\star P x_0.$$

In summary, the infinite-horizon cost $J(u)$ must be at least $x_0^\star P x_0$, while the *optimal* infinite-horizon cost $J(u_{opt})$ is at most $x_0^\star P x_0$. It follows that:

$$J(u_{opt}) = x_0^\star P x_0.$$

3. Since the system is LTI, it suffices to show that $\sigma(A - BR^{-1}B^\star P) \in \mathbb{C}^-$. Let $\lambda \in \sigma(A - BR^{-1}B^\star P)$ be arbitrarily given, with corresponding eigenvector $v \neq 0$. Then, from the ARE (6.2):

$$\begin{aligned} 0 &= v^\star \big[PA + A^\star P + Q - PBR^{-1}B^\star P\big] v \\ &= v^\star \big[P(A - BR^{-1}B^\star P) + (A - BR^{-1}B^\star P)^\star P + Q + PBR^{-1}B^\star P\big] v \\ &= 2Re(\lambda) \cdot v^\star P v + v^\star (Q + PBR^{-1}B^\star P) v \end{aligned}$$

Since $P \geq 0, Q > 0$, and $R \geq 0$, we have $Re(\lambda) \leq 0$.

Now, suppose by contradiction that $Re(\lambda) = 0$. Then, from the above chain of equalities, we have:

$$v^\star (Q + PBR^{-1}B^\star P) v = 0$$

. Recall that v was originally defined as an eigenvector of $(A - BR^{-1}B^\star P)$, with corresponding eigenvalue λ, and that $Q = C^\star C \geq 0$ and $R > 0$. Thus:

$$\begin{aligned} &\begin{cases} (A - BR^{-1}B^\star P)v = \lambda v, \\ Cv = 0, \\ B^\star P v = 0 \end{cases} && \Rightarrow \begin{cases} Av &= \lambda v, \\ Cv &= 0, \end{cases} \\ &\Rightarrow \begin{bmatrix} \lambda I - A \\ C \end{bmatrix} v = 0 \end{aligned}$$

Since (A, C) is observable, we have $v = 0$ by the PBH test, in contradiction of the fact that $v \neq 0$ from its definition as an eigenvector of $(A - BR^{-1}B^\star P)$.

4. Again, we separate the proof into two claims.

- $\underline{P > 0}$:

  We have already established that $P \geq 0$. Now, let an initial state $x_0 \in \mathbb{R}^n$ be given such that $x_0^\star P x_0 = 0$. By the above statements in this theorem, $x_0^\star P x_0$ is, in fact, the infinite-horizon optimal cost, so this implies that:

$$\begin{aligned} 0 = J(u_{opt}) &= \int_0^\infty \left[x^\star C^\star C x + u_{opt}^\star R u_{opt}\right] dt = 0, \\ \Rightarrow y = Cx = 0, &\qquad u_{opt} = 0, \end{aligned}$$

since $R > 0$. In other words, if the observable system (A, B) is propagated from the initial state x_0 with no control, its output is identically 0, at all times $t \geq 0$. This implies that the state $x(t)$ itself must be identically 0, at all times $t \geq 0$. In particular, $x_0 = 0$. This establishes $P > 0$, as desired.

- P is the unique positive semi-definite solution to the ARE (6.2)

  Suppose there exists another $P' \geq 0$ satisfying the ARE (6.2). Fix any initial state x_0, and consider the new finite-horizon cost function:

$$J'^{t_1}(u_{[0,t_1]}) \equiv \int_0^{t_1} \left[x^\star Q x + u^\star R u\right] dt + x^\star(t_1)\overline{P}x(t_1)$$

  From the solution to the finite-horizon LQR problem, we know that the optimal cost with respect to the above cost function is $x_0^\star P(0; \overline{P}, t_1)x_0$, where $P(0; \overline{P}, t_1)$ denotes a positive semi-definite satisfying the RDE:

$$\dot{P} = -PA - A^\star P - Q + PBR^{-1}B^\star P, \qquad P(t_1) = \overline{P}.$$

  However, by definition, $\overline{P}$ itself satisfies the ARE:

$$0 = -PA - A^\star P - Q + PBR^{-1}B^\star P$$

  Thus, $P(0; \overline{P}, t_1) = \overline{P}$, and so the optimal cost is $x_0^\star \overline{P} x_0$. Since this holds for any arbitrary t_1, we conclude that this is also the infinite-horizon optimal cost.

  On the other hand, we can directly calculate an expression for the infinite-horizon cost, as follows:

$$\begin{aligned}
\because J'^{t_1}(u_{[0,t_1]}) &\equiv \int_0^{t_1} \left[x^\star Q x + u^\star R u\right] dt + x^\star(t_1)\overline{P}x(t_1), \\
\Rightarrow J'^{\infty}(u_{[0,\infty)}) &= \int_0^{\infty} \left[x^\star Q x + u^\star R u\right] dt + \lim_{t_1 \to \infty} x^\star(t_1)\overline{P}x(t_1) \\
&= \int_0^{\infty} \left[x^\star Q x + u^\star R u\right] dt \geq \int_0^{\infty} \left[x^\star Q x + u_{opt}^\star R u_{opt}\right] dt \\
&= J^{\infty}(u_{opt}) = x_0^\star P x_0
\end{aligned}$$

  where the last two equalities follow from the fact that $\dot{x_{opt}} = (A - BR^{-1}B^\star P)x_{opt}$ is exponentially stable, and so $x(t_1) \to 0$ as $t_1 \to \infty$. Now, equality holds when we take the control u to be $u_{opt} = -R^{-1}B^\star P x$. The corresponding optimal cost is thus $x_0^\star P x_0$.

  In summary, the infinite-horizon optimal cost is $x_0^\star \overline{P} x_0 = x_0 P x_0$ for any initial state x_0. Taking x_0 to be $e_i, e_j, e_i + e_j$, and repeating the logic used to prove the boundedness of $x_0^\star P(0, t_1)x_0$ for any x_0 and $t_1 \geq 0$, we find that $\overline{P} = P$.

■

Remark.

1. If (A, B) were not controllable, the finite-horizon optimal cost $x_0^\star P(0, t_1)x_0$ may be unbounded above as $t_1 \to \infty$. For example, if $\dot{x} = ax$, where $a > 0$, then $x(t) = e^{at}$ is clearly unbounded above, and so is the infinite-horizon optimal cost:

$$J(u_{[0,t_1]}) = \int_0^{t_1} \left[x^\star Q x + u^\star R u\right] dt$$

   regardless of our choice of input (since $B = 0$ in this case). We reiterate that the reason for this phenomenon is that (A, B) is uncontrollable.

2. In fact, the optimal control $u_{opt}(t) = -R^{-1}B^\star P(t, t_1)x(t)$ given above is unique. The proof of this statement follows by applying the Hamilton-Jacobi-Bellman equation (see Appendix).

Examples

Example (**Finite-Horizon Optimal Cost**). Consider the scalar integrator and associated cost:

$$\dot{x} = u,\ x(0) = x_0,$$
$$J(u) = \int_0^{t_1} \left[x^2 + u^2\right] dt$$

for any $t_1 \geq 0$. Suppose we wish to minimize the cost for some fixed t_1. From the above equations, we have $A = 0, B = 1, Q = 1, R = 1, M = 0$. Substituting into the RDE (6.1) and solving for P, we have:

$$\begin{aligned}
&\dot{P} = P^2 - 1, \qquad P(t_1) = 0,\\
\Rightarrow &\int_{P(t)}^{P(t_1)} \frac{1}{P^2 - 1} dP = \int_t^{t_1} d\tau,\\
\Rightarrow &-\tanh^{-1}(P(t_1)) + \tanh^{-1}(P(t)) = t_1 - t,\\
\Rightarrow &P(t) = \tanh(t_1 - t),
\end{aligned}$$

since $P(t_1) = 0$. Substituting back into our expression for $u(t)$, we have:

$$u(t) = -\tanh(t_1 - t)x(t)$$

as the optimal control.

Remark. It is interesting to observe that, if we had defined $R = -1$ instead, the RDE (6.1)

$$\begin{aligned}
&\dot{P} = -P^2 - 1, \qquad P(t_1) = 0,\\
\Rightarrow\ &P(t) = \tan(t - t_1),
\end{aligned}$$

with corresponding solution:

and control $u(t) = -\tan(t - t_1)x$, which is not well-defined when $t - t_1$ is an odd multiple of $\pi/2$.

This illustrates the importance of postulating that $R \geq 0$.

The next example uses a simple, scalar system to illustrate that, without assuming (A, C) observable, we cannot guarantee that the input achieving the infinite-horizon optimal cost is stabilizing.

Example (**Infinite-Horizon Optimal Cost**). Consider the following system and associated cost:

$$\begin{aligned} \dot{x} &= x + u,\ x(0) = x_0, \\ J(u) &= \int_0^\infty u^2(t)\,dt \end{aligned}$$

Suppose we to calculate the finite-horizon optimal control and cost for the system. From the above equations, we have $A = 1, B = 1, Q = 0, R = 1$. Substituting into the ARE (6.2), we have:

$$\begin{aligned} &\because PA + A^\star P + Q - PBR^{-1}B^\star P = O, \\ &\Rightarrow 2P - P^2 = O. \end{aligned}$$

Thus, $P = 0$ or 2. Let us consider the cost associated with each case:

- If $P = 0$, the corresponding control, closed-loop dynamics, and cost are:

$$\begin{aligned} u_{opt} &= 0, \\ \dot{x_{opt}} &= x_{opt}, \\ J(u_{opt}) &= 0. \end{aligned}$$

  respectively. Thus, the corresponding control optimizes the infinite-horizon cost, but fails to create a stable closed-loop system.

- If $P = 2$, the corresponding control and cost are:

$$\begin{aligned} u_{opt} &= -2x_{opt}, \\ \dot{x_{opt}} &= -x_{opt}, \\ J(u_{opt}) &= 4x_0^2. \end{aligned}$$

  respectively. Thus, the corresponding control generates an stable closed-loop system, but fails to optimize the infinite-horizon cost.

Finite-Horizon LQR Problem—Hamiltonian Method

Below, we present an alternative, but equivalent, formulation of the solution to the LQR problem. This method can be found on Lecture 11, pgs. 1-12 of Professor Claire Tomlin's Lecture Notes, as well as Chapter 2, pgs. 29-37 of the Callier & Desoer text desoer1. First, we define a slightly more specific version of the LQR optimal control problem.

Definition 6.5 (**Finite-Horizon LQR Problem** (Modified version))**.** *The* ***finite-horizon linear quadratic optimal control problem*** *(discussed in this subseciton) is defined as follows—Given the system:*

$$\begin{aligned} \dot{x} &= Ax + Bu, \qquad x(0) = x_0, \\ y &= x, \end{aligned}$$

evolving in the time interval $[0, t_1]$ *with* $x \in \mathbb{R}^n$*,* $u \in \mathbb{R}^{n_i}$*, and* (A, B) *controllable,* (A, C) *observable, find the input function* $u(\cdot) : [0, t_1]$ *that minimizes a* ***quadratic cost functional****:*

$$J\big(u(\cdot)\big) = \int_0^{t_1} \big[x^\star C^\star C x + u^\star u\big] dt + x(t_1)^\star S x(t_1) \tag{6.3}$$

where $C(\cdot) : \mathbb{R}^+ \to \mathbb{R}^{n_0}$ *is a piecewise continuous function,* $S \geq 0$*, and* $J(\cdot)$ *maps a continuous function with domain* $[0, t_1]$ *and codomain* $\mathbb{R}^{n_i}$ *(namely,* $u_{[0,t_1]}$ *into* $\mathbb{R}$*. (As with before, the argument* t *is abbreviated in the integrand, for ease of notation).*

Before continuing, we will review our definition of the *adjoint system* (or *dual system*), as given by (5.15), reproduced below.

Definition 6.6 (**Adjoint System** (Dual System))**.** *The* ***adjoint system*** *(or dual system) of the linear time-varying system:*

$$\Sigma : \begin{cases} \dot{x}(t) = A(t)x(t) + B(t)u(t) \\ y(t) = C(t)x(t) + D(t)u(t) \end{cases}$$

is defined as:

$$\overline{\Sigma} : \begin{cases} \dot{\overline{x}}(t) = -A^\star(t)\overline{x}(t) - C^\star(t)\overline{u}(t) \\ \overline{y}(t) = B^\star(t)\overline{x}(t) + D^\star(t)\overline{u}(t) \end{cases}$$

The following useful lemma follows. Its proof can be found on pg. 36 of the Callier & Desoer text [4].

Lemma 6.7 (**Pairing Lemma**)**.** *Let* $\overline{\Sigma}$ *be the dual system of* Σ*, with notations for inputs, states, and outputs given as defined above. Then:*

$$\langle \overline{x}(t), x(t) \rangle + \int_0^t \langle \overline{u}(\tau), y(\tau) \rangle \, d\tau = \langle \overline{x}(0), x(0) \rangle + \int_0^t \langle \overline{y}(\tau), u(\tau) \rangle \, d\tau$$

for any $t, t_0 \in \mathbb{R}^+$*,* $x(t_0), \tilde{x}(t) \in \mathbb{R}^n$*,* $u(\cdot) \in \mathcal{U}$*, and* $\overline{u}(\cdot) \in \overline{\mathcal{U}}$*.*

Remark. The remarkable versatility of the Pairing Lemma lies in the fact that at it holds regardless of our choice of $u(\cdot), \overline{u}(\cdot), x(0), \overline{x}(0)$, and thus grants us the ability to freely define these parameters.

Proof. By duality, we have:

$$\begin{aligned}\dot{\overline{x}} &= -A^\star \overline{x} - C^\star u,\\ \overline{y} &= B^\star \overline{x} + D^\star \overline{u}.\end{aligned}$$

Rearranging terms and combining the above inequalities, we have:

$$\begin{aligned}0 &= \langle \dot{\overline{x}} + A^\star \overline{x} + C^\star \overline{u}, x\rangle + \langle -\overline{y} + B^\star \overline{x} + D^\star \overline{u}, u\rangle\\ &= \langle \dot{\overline{x}}, x\rangle + \langle \overline{x}, Ax\rangle + \langle \overline{u}, Cx\rangle - \langle \overline{y}, u\rangle + \langle \overline{x}, Bu\rangle + \langle \overline{u}, Du\rangle\\ &= \langle \dot{\overline{x}}, x\rangle + \langle \overline{x}, Ax + Bu\rangle + \langle \overline{u}, Cx + Du\rangle - \langle \overline{y}, u\rangle\\ &= \langle \dot{\overline{x}}, x\rangle + \langle \overline{x}, \dot{x}\rangle + \langle \overline{u}, y\rangle - \langle \overline{y}, u\rangle\\ &= \frac{d}{dt}\langle \overline{x}, x\rangle + \langle \overline{u}, y\rangle - \langle \overline{y}, u\rangle\end{aligned}$$

We recover the desired result by rearranging terms and integrating from 0 to t. ■

In this alternative solution method, instead of deriving of the optimal input $u_{[0,t_1]}$ and cost $J(u_{[0,t_1]})$ directly and invoking properties of the Riccati differential equation, we will instead observe what happens to the cost functional $J(u(\cdot))$ when the input function u is perturbed by an infinitesimal amount, i.e. $u \to u + \delta u$.

Definition 6.8 (**Global Minimizer of the Cost Functional**). *Given a cost functional $J(u_{\text{opt}}(\cdot))$, a **global minimizer** of $J(u(\cdot))$ is an input function u_{opt} such that, for each $\epsilon > 0$ and piecewise continuous $\delta u : [0, t_1] \to \mathbb{R}^{n_i}$:*

$$J(u_{\text{opt}} + \epsilon\,\delta u) \geq J(u_{\text{opt}})$$

Proposition 6.9. *Consider the expansion of $J(u + \epsilon\,\delta u)$ (with $\epsilon > 0$) about $\epsilon = 0$, as expressed in the following form:*

$$J(u + \epsilon\,\delta u) = J(u) + \epsilon \cdot \delta J(\delta u) + o(\epsilon)$$

Then, for a piecewise continuous input $\overline{u}_{0,t_1]}$ to be a global minimizer of the cost functional J, it is necessary and sufficient that, for <u>any</u> piecewise continuous input perturbation $\delta u_{[0,t_1]}$:

$$\delta J(\delta u) = 0$$

in the expansion of $J(u)$ about u_{opt}.

Proof. Expanding $J(u + \epsilon\,\delta u)$, we have:

$$\begin{aligned}J(u + \epsilon\,\delta u) &= \int_0^{t_1} \left[\|u + \epsilon\,\delta u\|^2 + \|C(x + \epsilon\,\delta x)\|^2\right] dt + (x + \epsilon\,\delta x)^\star(t_1) S(x + \epsilon\,\delta x)(t_1)\\ &= J(u) + \epsilon \cdot \left[2\int_0^{t_1} \left(\langle u, \delta u\rangle + \langle Cx, C\delta x\rangle\right) dt + \langle Sx(t_1), \delta x(t_1)\rangle\right] + \epsilon^2 \cdot J(\delta u)\\ &\equiv J(u) + \epsilon \cdot \delta J(\delta u) + \epsilon^2 \cdot J(\delta u) \qquad (6.4)\end{aligned}$$

where we have defined:

$$\delta J(\delta u) = 2\int_0^{t_1} \big(\langle u, \delta u\rangle + \langle Cx, C\delta x\rangle\big)\, dt + \langle Sx(t_1), \delta x(t_1). \tag{6.5}$$

Observe that $\epsilon^2 \cdot J(\delta u)$ can also be expressed as $o(\epsilon)$.

" $\Rightarrow$ " : If $u_{[0,t_1]}$ is a global minimizer, then, for each $\epsilon \in \mathbb{R}$, we have:

$$0 \leq J(u + \epsilon\,\delta u) - J(u) = \epsilon \left[\delta J(\delta u) + \frac{o(\epsilon)}{\epsilon}\right]$$

Thus, for $\epsilon \to 0^+$, we have $\delta J(\delta u) \geq 0$; for $\epsilon \to 0^-$, we have $\delta J(\delta u) \leq 0$. We conclude that $\delta J(\delta u) = 0$.

" $\Leftarrow$ " : Suppose $\delta J(\delta u) = 0$. Then, from (6.4), we have:

$$J(u_{opt} + \epsilon\,\delta u) = J(u_{opt}) + \epsilon^2 \cdot J(\delta u) \geq J(u),$$

i.e. u_{opt} is minimizing.

■

Theorem 6.10 (**Finite-Horizon LQR Solution: Hamiltonian Method**). *Consider the $2n$-dimensional two point boundary value problem, given by:*

$$\begin{aligned} \begin{bmatrix}\dot{x}\\ \dot{\overline{x}}\end{bmatrix} &= \begin{bmatrix} A & -BB^\star \\ -C^\star C & -A^\star \end{bmatrix}\begin{bmatrix} x \\ \overline{x}\end{bmatrix}, \\ x(0) &= x_0, \\ \overline{x}(t_1) &= Sx(t_1), \end{aligned} \tag{6.6}$$

with corresponding matrix equation:

$$\begin{aligned} \begin{bmatrix}\dot{X}\\ \dot{\overline{X}}\end{bmatrix} &= \begin{bmatrix} A & -BB^\star \\ -C^\star C & -A^\star \end{bmatrix}\begin{bmatrix} X \\ \overline{X}\end{bmatrix}, \\ X(0) &= I, \\ \overline{X}(t_1) &= S. \end{aligned} \tag{6.7}$$

where the argument t is hidden in $A, B, C, X, \overline{X}$, for convenience of notation. Then $X(t)$ is invertible for each $t \in [0, t_1]$, and the optimal control to the finite-horizon LQR problem is the linear time-varying control:

$$u_{\text{opt}}(t) = -B^\star \overline{x}(t) \tag{6.8}$$

$$= -B^\star \overline{X} X^{-1} x(t) \tag{6.9}$$

with corresponding optimal cost:

$$J(u_{\text{opt}}(\cdot)) = \langle \overline{x}_0, x_0\rangle \tag{6.10}$$

$$= \langle \overline{X}(0)X^{-1}(0)x_0, x_0\rangle \tag{6.11}$$

Proof. We will divide the proof of this theorem into the following three components.

- $u_{opt}(t) = -B^\star \overline{x}(t)$ is the optimal control, with corresponding optimal cost $J(u_{opt}(\cdot))$.
- $X(t)$ is invertible for each $t \in [0, t_1]$.
- $\overline{x}(t) = \overline{X} X^{-1} x(t)$.

Notice that the first claim establishes (6.8) and (6.10), whereas the second and third establish (6.9) and (6.11). We begin the proof below.

- $u_{\text{opt}}(t) = -B^\star \overline{x}(t)$ is the optimal control, with corresponding optimal cost $J(u_{\text{opt}}(\cdot))$:

  If a perturbation δu in the input u results in a perturbation δx in the resulting trajectory x, the resulting trajectory is $\dot{x} + \delta\dot{x} = A(x + \delta x) + B(u + \delta u)$. Thus, the error system is:

$$\begin{aligned} \delta\dot{x} &= A\,\delta x + B\,\delta u, \qquad x(0) = x_0. \\ \delta\, y &= \delta\, x \end{aligned} \tag{6.12}$$

  By the definition of dual system (Theorem 5.15), the dual system of (6.12) is as follows:

$$\begin{aligned} \dot{\overline{x}} &= -A^\star \overline{x} - \overline{u}, \qquad \overline{x}(t_1) = \overline{x_1} \\ \overline{y} &= B^\star \overline{x} \end{aligned} \tag{6.13}$$

  Recall that this holds for *any* choice of $\overline{x_1}$ and $\overline{u}(t)$.

  Applying the Pairing Lemma (Lemma 6.7) to (6.12), (6.13), we have:

$$\langle \overline{x_1}, \delta x(t_1) \rangle + \int_0^{t_1} \langle \overline{u}, \delta x \rangle\, dt = \int_0^{t_1} \langle B^\star \overline{x}, \delta u \rangle\, dt$$

  Meanwhile, (6.5) gives the expression for the cost perturbation $\delta J(\delta u)$. Combining this with the above equation from the Pairing Lemma, we have:

$$\begin{aligned} \delta J(\delta u) &= \int_0^{t_1} \langle u, \delta u \rangle + \langle C^\star C x, \delta x \rangle\, dt + \langle S x(t_1), \delta x(t_1) \rangle \\ &= \int_0^{t_1} \langle u + B^\star \overline{x}, \delta u \rangle + \langle C^\star C x - \overline{u}, \delta x \rangle\, dt + \langle S x(t_1) - \overline{x_1}, \delta x(t_1) \rangle \end{aligned}$$

  Again, this holds for *any* choice of $\overline{x_1}$ and $\overline{u}(t)$. Now, Proposition 6.9 impliess that, to minimize $\delta J(\delta u)$, we wish to find a choice of u such that $\delta J(\delta u) = 0$. The form of the above equation implies that this is easiest when we choose:

$$\begin{aligned} \overline{u} &= C^\star C x, \\ \overline{x_1} &= S\, x(t_1), \end{aligned}$$

which rids us of the second and third terms (those unrelated to δu). With this choice, $\delta J(\delta u)$ now becomes:

$$\delta J(\delta u) = \int_0^{t_1} \langle u + B^\star \overline{x}, \delta u \rangle + \langle C^\star C x - \overline{u}, \delta x \rangle \, dt$$

To ensure that this value stays 0 for any choice of δu, we must choose:

$$u_{\text{opt}} = -B^\star \overline{x}$$

for our control. This verifies that the optimal control is given by (6.8). The corresponding optimal cost is thus:

$$\begin{aligned} J(u_{\text{opt}}) &\equiv \int_0^{t_1} \left[\langle u, u \rangle + \langle Cx, Cx \rangle \right] dt + x^\star(t_1) S \, x(t_1) \\ &= \int_0^{t_1} \left[\langle -B^\star x, u \rangle^2 + \langle \overline{u}, x \rangle \right] dt + x^\star(t_1) S \, x(t_1) \\ &= \langle \overline{x}_0, x_0 \rangle \end{aligned}$$

where the final equality follows from the Pairing Lemma. This verifies (6.9).

- $X(t)$ is invertible for each $t \in [0, t_1]$:

  Consider (6.7). Suppose by contradiction that there exists some $\tau \in [0, t_1)$ such that $X(\tau)$ is singular (since $X(t_1) = I$, we have $\tau \neq t_1$). This implies there exists some $k \neq 0$ such that $X(\tau)k = 0$. Now, observe that:

$$\begin{bmatrix} x(t) \\ \overline{x}(t) \end{bmatrix} = \begin{bmatrix} X(t) \\ \overline{X}(t) \end{bmatrix} k$$

  solves the differential equations in (6.6) for each $t \in [\tau, t_1]$, with boundary conditions changed to:

$$\begin{aligned} x(\tau) &= X(\tau)k = 0, \\ \overline{x}(t_1) &= \overline{X}(t_1)k = Sk. \end{aligned}$$

  Substituting into (6.9), we have:

$$\begin{aligned} 0 = \langle \overline{x}(\tau), x(\tau) \rangle &= J(u_{[\tau, t_1]}) \\ &= \int_\tau^{t_1} \left[\|u\|^2 + \|Cx\|^2 \right] dt + x^\star(t_1) S x(t_1). \end{aligned}$$

  Since $S > 0$, each of the above terms in the final integral must be zero, i.e. $x(t_1) = 0$, and $u(t) = 0$, $Cx(t) = 0$ for each $t \in [\tau, t_1]$. Thus, the original dynamics become:

$$\dot{x}(t) = Ax(t) + Bu(t) = Ax(t), \qquad x(\tau) = 0,$$

  which implies $x(t_1) = 0$, contradicting the fact that, since $X(t_1)$ is invertible and $k \neq 0$:

$$x(t_1) = X(t_1)k$$

  is nonzero. The claim follows by contradiction.

- $\underline{\overline{x}(t) = \overline{X}X^{-1}x(t)}$:

  Our goal is to associate $\overline{x}(t)$ to $x(t)$. Now, observe that:

$$\begin{bmatrix} x(t) \\ \overline{x}(t) \end{bmatrix} = \begin{bmatrix} X(t) \\ \overline{X}(t) \end{bmatrix} k$$

  satisfy (6.7). The Fundamental Theorem of Differential Equations (Theorem 3.4) thus implies that they must be the unique solution to (6.7). Moreover, since $X(t)$ is non-singular for each $t \in [0, t_1]$, we have:

$$\overline{x}(t) = \overline{X}(t)X^{-1}(t)x(t),$$

  as claimed. Equations (6.10) and (6.11) thus follow. The proof is done.

■

Remark (**Comparison of the Two Approaches** ([4], Chapter 2, pg. 38)). It is interesting to compare the two approaches to the finite-horizon LQR problem. For the variant of the problem described in (6.5), the Riccati Differential Equation (RDE) approach gives the optimal control as:

$$u_{\text{opt}}(t) = -B^\star P x(t),$$

for each $t \in [0, t_1]$, with $P(t)$ as the unique positive definite solution to the (quadratic) RDE $\dot{P} + PA + A^\star P - PBB^\star P + C^\star C = 0$, $P(t_1) = S$, whereas the Hamiltonian method gives the optimal control as:

$$u_{\text{opt}}(t) = -B^\star \overline{X} X x(t),$$

for each $t \in [0, t_1]$, with the tuple $X, \overline{X}$ satisfying the (linear) differential equations given by (6.7).

These two formulations are, in fact, equivalent, as can be seen by "establishing a bijection" between the solution to the RRDE (modified for this variant of the problem, as given above), and the solution to the Hamiltonian approach given by (6.6).

- Hamiltonian Approach $\rightarrow$ RDE:

  Let $X(t), \overline{X}(t)$ be the solution to the Hamiltonian formulation (6.7), and define $P_1 \equiv \overline{X}X^{-1}$. Then, from (6.7), we have:

$$\begin{aligned} \dot{P}_1 &= \dot{\overline{X}}X^{-1} - \overline{X}X^{-1}\dot{X}X^{-1} \\ &= (-C^\star CX - A^\star \overline{X})X^{-1} - P_1(AX - BB^\star \overline{X})X^{-1} \\ &= -C^\star C - A^\star P_1 - P_1 A + P_1 BB^\star P_1, \end{aligned}$$

  with $P_1(t_1) = \overline{X}(t_1)X^{-1}(t_1) = S$. Thus, $P_1 \equiv \overline{X}X^{-1}$ is the unique solution to the RDE $\dot{P} + PA + A^\star P - PBB^\star P + C^\star C = 0$.

Now, since $P_1(t_1) = S$ is positive definite, and therefore Hermitian, so is $\dot{P}_1$. Thus, $P_1(t)$ is Hermitian for each $[0, t_1]$. Meanwhile, from (6.11), we have:

$$\begin{aligned} x_0^\star P(t_0) x_0 &= x_0^\star \overline{X}(t_0) X^{-1}(t_0) x_0 = \langle \overline{x}(t_0), x(t_0) \rangle^\star \\ &= J(u_{\text{opt},[t_0,t]])^\star \geq 0, \end{aligned}$$

for any choice of initial time t_0, which implies that each $P(t)$ is positive semidefinite. Since RDEs yield unique positive semidefinite solutions, $P(t)$ must the *unique* positive definite solution to $\dot{P} + PA + A^\star P - PBB^\star P + C^\star C = 0$, $P(t_1) = S$.

- RDE $\to$ Hamiltonian Approach:

  Conversely, suppose P is the unique positive definite solution to the RDE $\dot{P} + PA + A^\star P - PBB^\star P + C^\star C = 0$, $P(t_1) = S$. Following the RDE formulation, we know that the optimal control is $u(t) = -B^\star P x(t)$. Thus, we have:

$$\begin{aligned} \begin{bmatrix} \dot{X} \\ (\dot{PX}) \end{bmatrix} &= \begin{bmatrix} \dot{X} \\ \dot{P}X + P\dot{X} \end{bmatrix} \\ &= \begin{bmatrix} (A - BB^\star P)X \\ (-PA - A^\star P + PBB^\star P - C^\star C)X + P(A - BB^\star P)X \end{bmatrix} \\ &= \begin{bmatrix} (A - BB^\star P)X \\ (-A^\star P - C^\star C)X \end{bmatrix} \\ &= \begin{bmatrix} A & -BB^\star \\ -C^\star C & -A^\star \end{bmatrix} \begin{bmatrix} X \\ PX \end{bmatrix}. \end{aligned}$$

  Thus, the tuple (X, PX) satisfies the Hamiltonian formulation (6.7). Uniqueness follows from the Fundamental Theorem (Theorem 3.4).

Although the two approaches are equivalent, it is interesting to contemplate their differences. The RDE $\dot{P} + PA + A^\star P - PBB^\star P + C^\star C = 0$, $P(t_1) = S$ is a quadratic differential equation, whereas in the Hamiltonian formulation, (6.7) gives a system of two linear differential equations. In effect, the Hamiltonian approach swaps the quadratic RDE for two linear ODEs, at the cost of solving two variables $(X, \overline{X})$ instead of one (P).

Finally, we conclude this section with a specific example of an optimal control problem for a non-linear system.

Example. Consider the non-linear system:

$$\dot{x} = f(x, u, t), \qquad x(0) = x_0.$$

Given a cost functional $J(u(\cdot))$ *that depends only on the final state*, i.e. of the form:

$$\begin{aligned} J(u(\cdot)) &= g(x(t_1)) \\ &= g(s(t_1, 0, x_0, u(\cdot)), \end{aligned}$$

find the optimal piecewise continuous control that minimizes $J(u(\cdot))$.

Solution : In general, there is no closed-form solution. However, in practice, we can start from a reasonable guess, and try to improve on it. Let $u_i(\cdot), x_i(\cdot)$ be an initial guess, and perturb it as follows:

$$\begin{aligned} u_i(\cdot) &\to u_i(\cdot) + \delta u(\cdot), \\ x_i(\cdot) &\to x_i(\cdot) + \delta x(\cdot). \end{aligned}$$

The problem is now to find a perturbation $\delta u(\cdot)$ to decrease the cost, i.e. such that:

$$g\big(x_i(t_1) + \delta x(t_1)\big) < g(x_i(t_1))$$

Since $\dot{x} = f(x, u, t)$, the dynamics of the perturbation are:

$$\delta x = \underbrace{D_1 f(x_i, u_i, t)}_{\equiv A(t)} \delta x + \underbrace{D_2 f(x_i, u_i, t)}_{\equiv B(t)} \delta u, \qquad \delta x(0) = 0,$$

Consider the adjoint system given by:

$$\dot{\overline{x}} = -A^\star(t)\overline{x}(t), \qquad \overline{x}(t_1) = D_{x_1} g\Big|_{x_i(t_1)},$$

(which holds if we set $y = 0$ for the original system). Let $\Phi(t, \tau)$ be the state transition matrix describing the perturbation dynamics. Then $\overline{x}(t) = \Phi^\star(t, t_1)\overline{x}(t_1) = \Phi^\star(t, t_1) D_{x_1} g|^\star_{x_i(t_1)}$, and:

$$\begin{aligned} \delta g(t_1) &= D_{x_1} g|_{x_i(t_1)} \cdot \delta x(t_1) \\ &= \int_0^{t_1} D_{x_1} g|_{x_i(t_1)} \, \Phi(t, \tau) \, B(\tau) \, \delta u(\tau) \, d\tau \\ &= \int_0^{t_1} \overline{x}(\tau)^\star \, B(\tau) \, \delta u(\tau) \, d\tau. \end{aligned}$$

Observe that the input perturbation:

$$\begin{aligned} \delta u(t) &\equiv -\alpha B^\star \overline{x}(t) = -\alpha B^\star \overline{X} X^{-1} x(t) \\ &= -\alpha B^\star P x(t) \end{aligned}$$

always renders the cost perturbation $\delta g(t_1)$ non-positive. However, Proposition 6.9 implies that, to optimize the cost, it is necessary and sufficient to choose a control $u_i(\cdot)$ such that $\delta g(t_1) = 0$ for *any* input perturbation δu. Thus, we must require:

$$B^\star(t) P(t) = 0$$

as a *necesary* condition for reaching a local optimum. In other words, a necessary condition for reaching a local optimum is:

$$\begin{aligned} \dot{x}_i &= f(x_i, u_i, t), \qquad x_i(0) = x_0, \\ \dot{\overline{x}} &= -(D_1 f)(x_i, u_i, t)\overline{x}, \qquad \overline{x}(t_1) = D_1 g|_{x_i(t_1)} \end{aligned}$$

with $(D_2 f)(x_i, u_i, t)\overline{x}(t) = 0$.

Remark (**Hamilton's Equations**). Alternatively, if we define the Hamiltonian as:

$$H(x, u, \overline{x}, t) = \overline{x}^\star f(x, u, t),$$

the necessary conditions described above can be rephrased as follows—Choose a control input u_i such that $\frac{\partial H}{\partial u}\big|_{u_i} = 0$, and set:

$$\begin{aligned} \dot{x} &= \left(\frac{\partial H}{\partial \overline{x}}\right)^\star, \qquad x(0) = x_0, \\ \dot{\overline{x}} &= -\left(\frac{\partial H}{\partial x}\right)^\star, \qquad \overline{x}(t_1) = (D_1 g)^\star(x(t_1)). \end{aligned}$$

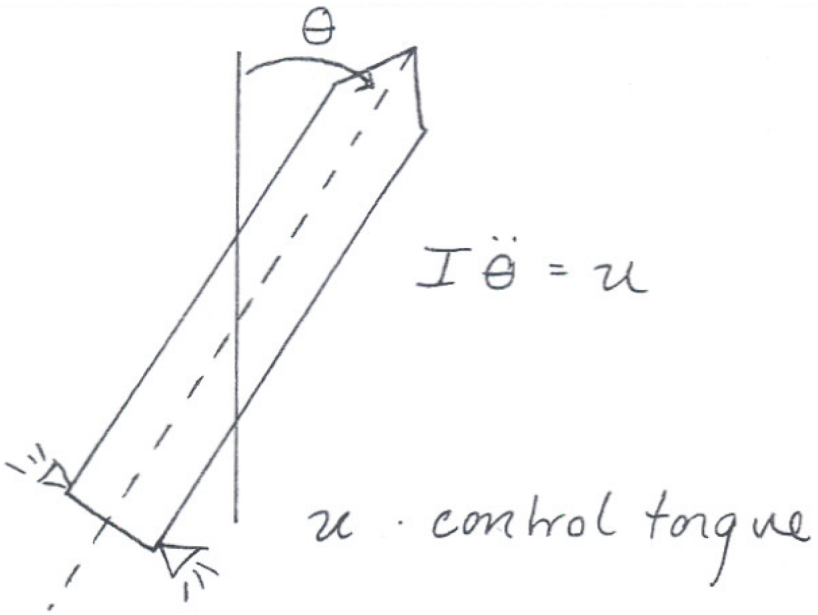
These are known as Hamilton's Equations. More details can be found in Chapter 2 of [6].

Example (**Optimal Control of a Single-Axis Satellite Altitude**). Suppose a given single-axis satellite, as shown below, can be approximately described by the time-invariant system:

$$\begin{aligned} \dot{x} &= Ax + Bu, \\ y &= x \end{aligned}$$

where A and B are given by:

$$A \equiv \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \qquad B \equiv \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$



Here, we wish to choose $P \geq 0, R > 0$ to minimize the infinite-horizon cost given by:

$$J \equiv \int_0^\infty (x^\star Q x + u^\star R u)\, dt$$

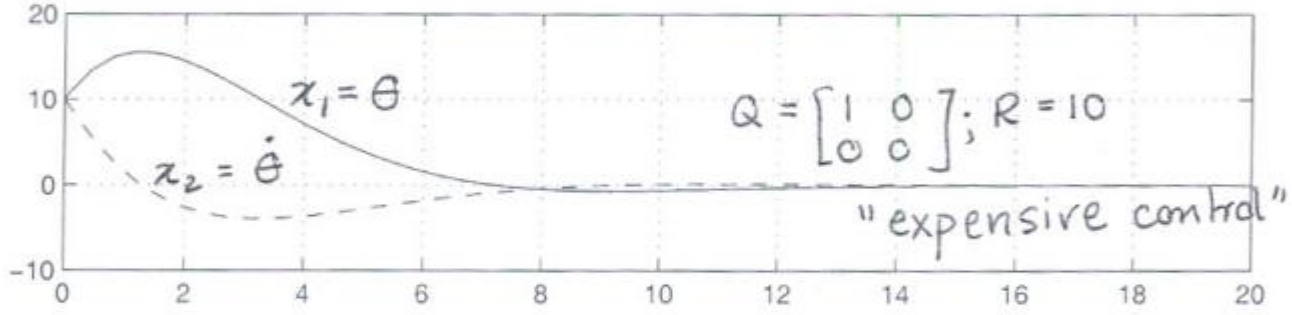
The figure on the next page shows MATLAB solutions with parameters set at:

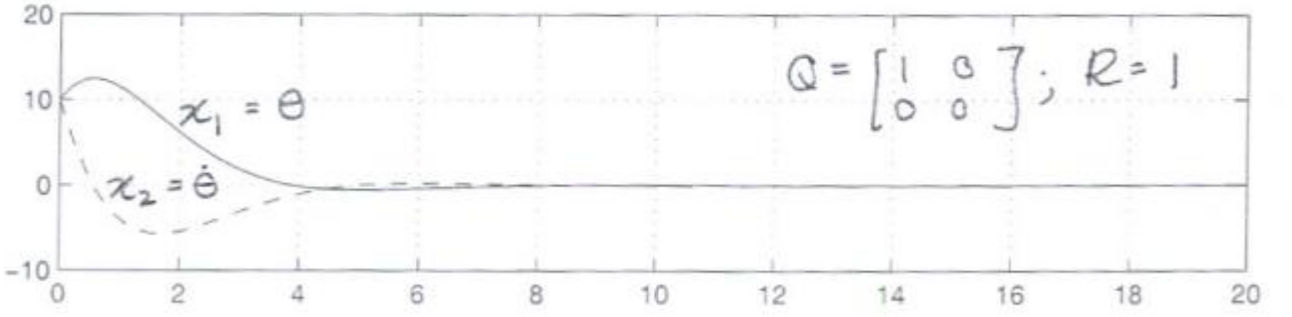
$$Q = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \qquad \text{and}$$
$$R = 10, 1, 0.01,$$

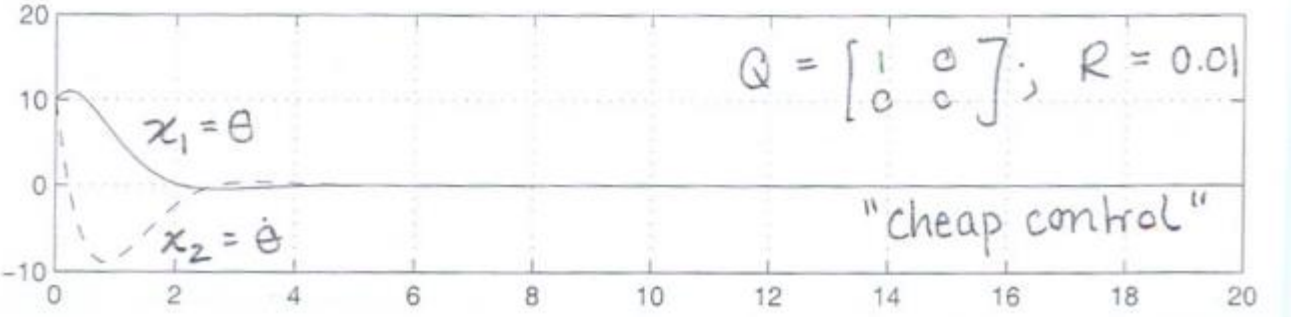
respectively.

```
%       LQR example
% call_satellite.m

global A;
global B;
global K;


A = [0 1; 0 0];
B = [0;1];
Q = [1 0;0 0];
R = 0.1;

[K,P,E] = lqr(A,B,Q,R);

x0 = [10 10];
t0 = 0; tf = 20;
[T,x]=ode23('satellite', [t0,tf], x0);

plot(T, x(:,1),T, x(:,2), '--');
```

```
%       LQR example
% satellite.m

function [xdot] = satellite(t,x)


global A;
global B;
global K;

xdot = (A-B*K)*x;
```







6.2 Hamilton-Jacobi-Bellman Equation

Dynamic Programming:

To motivate subsequent discussions regarding the Hamilton-Jacobi-Bellman Equation, we first explain the principles of dynamic programming as they apply to a finite-time, discrete system with a finite number of states and a finite number of control input choices at each state. For example, suppose the state space X and input space U consist of n and n_i elements, respectively, and that the dynamics of the system evolve according to:

$$x_{k+1} = f(x_k, u_k), k = \{0, 1, \cdots, T-1\}$$

where $T \in \mathbb{N}$, and $x_k \in X$, $u_k \in U$ for each $k = \{0, 1, \cdots, T\}$. Suppose we wish to minimize some cost function:

$$J(x_0, u_0, u_1, \cdots, u_{T-1}) + K(x_T),$$

where $J : X \times U^T \to \mathbb{R}$ and $K : X \to \mathbb{R}$ are the *running cost* and *terminal cost*, respectively.

The most straightforward method is to enumerate all possible state trajectories going forward from (x_0, t_0), and compare the resulting costs. An alternative, and more efficient, approach to the problem is to apply dynamic programming, in the form of *backwards induction*. At time T, the terminal cost is known for each state x_k. Now, at time $T-1$, find a state and control pair u_{T-1} that minimizes the cost-to-go, i.e. the sum of the one-step running cost from $T-1$ to T, and the terminal cost at time T. Take note of the corresponding one-step trajectories, and repeat this process for times $T-2, T-3, \cdots, 1, 0$, by working backwards using the computed costs-to-go as the terminal cost for each step. We claim that, at the end of this procedure, we will have obtained an optimal trajectory. This is because *the backward induction process eliminates all locally sub-optimal paths, which cannot be a part of any optimal trajectory.*

Let us compare the computational complexity of each approach. For the straightforward approach, T additions are required to compute the cost for each of roughly n_i^T possible trajectories, resulting in a computational cost of approximately $O(n_i^T T)$. For the dynamic programming approach, there are T points in time, each of which requires a backward induction step of choosing one of n possible states and one of n_i inputs, corresponding to a computational cost of approximately $O(n_i^T T)$. Thus, for a fixed number of states and inputs (i.e. fixed n, n_i), it is clear that dynamic programming is far more efficient than the forward search. This is because dynamic programming allows us to eliminate large subsets of trajectories that are globally sub-optimal in each unit time interval, by rejecting fragments of these sub-optimal paths at each point in time. A brute-force forward search algorithm cannot accomplish this, since a locally sub-optimal path fragment may be part of a globally optimal trajectory, and vice versa.

The advantages of using dynamic programming actually extend beyond ensuring higher computational efficiency. In fact, the backwards induction process identifies an optimal control for each initial state x_0. In other words, dynamic programming generates an optimal control in the form of a *state feedback*. Thus, this approach realizes the following quote given by Richard E. Bellman, who first introduced dynamic programming in 1953: "In place of determining the

optimal sequence of decisions from a fixed state of the system, we wish to determine the optimal decision to be made at any state of the system. Only if we know the latter, do we understand the intrinsic structure of the solution." [3]

Remark (**Curse of Dimensionality**). If n and n_i are large, the number of operations required to find the optimal trajectory will still grow rapidly, regardless of whether one applies the straightforward method or dynamic programming. This is known as "the curse of dimensionality," a somewhat nebulous description. As Professor Alessandro Astolfi likes to quip: "At what dimension does the problem become cursed?"

Principle of Optimality in Optimal Control:

Below, we apply dynamic programming to solve a class of optimal control problem that is more general than the LQR problem. In particular, given a continuous time system:

$$\dot{x} = f(x, u, t),$$

with $x \in \mathbb{R}^n, u \in \mathcal{U} \subset \mathbb{R}^{n_i}, t \geq 0$, and some well-behaved function $f : \mathbb{R}^n \times \mathcal{U} \times [0, \infty) \to \mathbb{R}^n$, we wish to minimize the global cost functional:

$$J(x_0, u) = \int_0^{t_1} L\big(\tau, x(\tau), u(\tau)\big)\, d\tau + K(x(t_1)),$$

where $L : \mathbb{R} \times \mathbb{R}^n \times \mathcal{U}$ and $K : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$ are called the *running cost* (or *Lagrangian*) and *terminal cost*, respectively.

Dynamic programming suggests that we should consider the cost-to-go at each $t \in [t_0, t_1]$:

$$J(t, x(t), u_{[t,t_1]}) = \int_t^{t_1} L\big(\tau, x(\tau), u(\tau)\big)\, d\tau + K(x(t_1)).$$

Our objective is to find the global minimum of this quantity:

$$\begin{aligned} V(t, x(t)) &\equiv \inf_{u_{[t_0,t_1]}} J(t, x(t), u_{[t,t_1]}) \\ &= \inf_{u_{[t_0,t_1]}} \left\{ \int_t^{t_1} L\big(\tau, x(\tau), u(\tau)\big)\, d\tau + K(x(t_1)) \right\} \end{aligned}$$

Below, we apply the principle of dynamic programming to $V(t, x(t))$.

Proposition 6.11 (**Principle of Optimality in Optimal Control**). *For every $x \in \mathbb{R}^n$, and every $t, \Delta t$ such that $0 \leq t < t + \Delta t < t_1$, the value function $V(t, x(t))$ defined above satisfies:*

$$V(t, x(t)) = \inf_{u_{[t,t+\Delta t]}} \left\{ \int_t^{t+\Delta t} L\big(\tau, x(\tau), u(\tau)\big)\, d\tau + V(t + \Delta t, x(t + \Delta t)) \right\} \tag{6.14}$$

where $x(\cdot)$ on the right-hand side denotes the state trajectory, starting at $x(t) = x$, that corresponds to the control $u_{[t,t+\Delta t]}$.

Proof. Let $\overline{V}(t, x(t))$ be the expression defined on the right-hand side, i.e.:

$$\overline{V}(t, x(t)) = \inf_{u_{[t,t+\Delta t]}} \left\{ \int_t^{t+\Delta t} L\big(\tau, x(\tau), u(\tau)\big)\, d\tau + V(t + \Delta t, x(t + \Delta t)) \right\}$$

Below, we aim to show that, for any $\epsilon > 0$:

$$\begin{aligned} V(t, x(t)) &< \overline{V}(t, x(t)) + \epsilon, \\ \overline{V}(t, x(t)) &< V(t, x) + \epsilon, \end{aligned}$$

which establishes $V(t, x(t)) = \overline{V}(t, x(t))$.

Let $\epsilon > 0$. By definition of $V(t, x(t))$, there exists some control $u_{\epsilon,[t,t_1]}$, with corresponding trajectory $x_{\epsilon,[t,t_1]}$, such that:

$$\int_t^{t_1} L\big(\tau, x_\epsilon(\tau), u_\epsilon(\tau)\big)\, d\tau + K(x(t_1)) < V(t, x(t)) + \epsilon$$

Then, by definition of $\overline{V}(t, x(t))$ and $V(t + \Delta t, x_\epsilon(t + \Delta t)$, we have:

$$\begin{aligned} \overline{V}(t, x(t)) &\leq \int_t^{t_1} L\big(\tau, x_\epsilon(\tau), u_\epsilon(\tau)\big)\, d\tau + V(t + \Delta t, x_\epsilon(t + \Delta t)) \\ &\leq \int_t^{t+\Delta t} L\big(\tau, x_\epsilon(\tau), u_\epsilon(\tau)\big)\, d\tau + \int_{t+\Delta t}^{t_1} L\big(\tau, x_\epsilon(\tau), u_\epsilon(\tau)\big)\, d\tau \\ &\qquad + V(t + \Delta t, x_\epsilon(t + \Delta t)) \\ &\leq \int_t^{t_1} L\big(\tau, x_\epsilon(\tau), u_\epsilon(\tau)\big)\, d\tau + K(x(t_1)) \\ &< V(t, x(t)) + \epsilon \end{aligned}$$

Similarly, there must exist controls $u_{\epsilon_1,[t,t+\Delta t]}$ and $u_{\epsilon_2,[t+\Delta t,t_1]}$ such that:

$$\begin{aligned} \int_t^{t+\Delta t} L\big(\tau, x(\tau), u_{\epsilon_1}(\tau)\big) + V(t + \Delta t, x(t + \Delta t)) &< \overline{V}(t, x(t)) + \frac{1}{2}\epsilon, \\ \int_{t+\Delta t}^{t_1} L\big(\tau, x(\tau), u_{\epsilon_2}(\tau)\big) + K(x(t_1)) &< V(t + \Delta t, x(t + \Delta t)) + \frac{1}{2}\epsilon. \end{aligned}$$

Let $u_{\epsilon,[t,t_1]}$ be defined such that $u_{\epsilon,[t,t_1]}(\tau) = u_{\epsilon_1,[t,t+\Delta t]}(\tau)$ when $\tau \in [t, t + \Delta t)$, and $u_{\epsilon,[t,t_1]}(\tau) = u_{\epsilon_2,[t+\Delta t,t_1]}(\tau)$ when $\tau \in [t + \Delta t, t_1)$. Thus:

$$\begin{aligned} V(t, x) &\leq \int_t^{t_1} L\big(\tau, x(\tau), u_\epsilon(\tau)\big)\, d\tau + K(x(t_1)) \\ &\leq \int_t^{t+\Delta t} L\big(\tau, x(\tau), u_{\epsilon_1}(\tau)\big)\, d\tau + \int_{t+\Delta t}^{t_1} L\big(\tau, x(\tau), u_{\epsilon_2}(\tau)\big)\, d\tau + K(x(t_1)) \\ &< \int_t^{t+\Delta t} L\big(\tau, x(\tau), u_\epsilon(\tau)\big)\, d\tau + V(t + \Delta t, x(t + \Delta t)) + \frac{1}{2}\epsilon \\ &< \overline{V}(t, x) + \epsilon \end{aligned}$$

In summary, we have $\overline{V}(t,x) < V(t,x) + \epsilon$ and $V(t,x) < \overline{V}(t,x) + \epsilon$, so $V(t,x) = \overline{V}(t,x)$, as claimed. The proof is done.

■

Remark. In the previous section, we have essentially described the principle of optimality in the context of controlling a n-state, n_i-input, finite-horizon discrete-time system. In the definition directly above, the context became the optimal control of continuous systems. However, Bellman originally defined the principle of optimality as follows, to cover a much broader scope of problems— "An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision." [3]

Hamilton-Jacobi-Bellman Equation:

The Hamilton-Jacobi-Bellman (HJB) Equation can be thought of as the differential analogue of the principle of optimality. That is, whereas the principle of optimality encapsulates the spirit of dynamic programming in a difference equation, the HJB Equation describes the essence of dynamic programming using a differential equation. It is derived by taking $\Delta t \to 0$ in the principle of optimality.

Theorem 6.12 (**Hamilton-Jacobi-Bellman Equation (HJB)**). *For each $x \in \mathbb{R}^n$ and $t \in [t_0, t_1)$, the **Hamilton-Jacobi-Bellman (HJB) Equation**:*

$$-\frac{\partial V}{\partial t}(t, x(t)) = \inf_{u \in \mathcal{U}} \left\{ L(t, x(t), u(t)) + \left\langle \frac{\partial V}{\partial x}(t, x(t)), f(t, x, u) \right\rangle \right\} \tag{6.15}$$

holds. It is the differential analogue of the principle of optimality.

Proof. The principle of optimality, (6.14), states that:

$$V(t, x(t)) = \inf_{u_{[t,t+\Delta t]}} \left\{ \int_t^{t+\Delta t} L\big(\tau, x(\tau), u(\tau)\big)\, d\tau + V(t + \Delta t, x(t + \Delta t)) \right\}.$$

When $\Delta t \to 0$, the two terms on the right-hand side of the above expression become:

$$\int_t^{t+\Delta t} L\big(\tau, x(\tau), u(\tau)\big)\, d\tau = L\big(t, x(t), u(t)\big)\Delta t + o(\Delta t)$$

and:

$$\begin{aligned} &V(t + \Delta t, x(t + \Delta t)) \\ =&V(t, x(t)) + \frac{\partial V}{\partial t}(t, x(t))\Delta t + \left\langle \frac{\partial V}{\partial x}(t, x(t)) \cdot f(t, x, u) \right\rangle \Delta t + o(\Delta t) \end{aligned}$$

where $V(t + \Delta t, x(t + \Delta t))$ has been expanded using the Chain Rule:

$$\frac{dV}{dt} = \frac{\partial V}{\partial t} + \frac{\partial V}{\partial x} \cdot \frac{dx}{dt} = \frac{\partial V}{\partial t} + \frac{\partial V}{\partial x} \cdot f(t, x, u)$$

After substituting back into the principle of optimality, (6.14), canceling out the term $V(t, x(t))$ on both sides, and dividing by Δt, we have:

$$V(t, x(t)) = \inf_{u_{[t,t+\Delta t]}} \left\{ L\big(t, x(t), u(t)\big)\Delta t + V(t, x(t)) + \frac{\partial V}{\partial t}(t, x(t))\Delta t + \left\langle \frac{\partial V}{\partial x}(t, x(t)) \cdot f(t, x, u) \right\rangle \Delta t \right\}$$

$$\Rightarrow -\frac{\partial V}{\partial t}(t, x(t)) = \inf_{u_{[t,t+\Delta t]}} \left\{ L\big(t, x(t), u(t)\big) + \left\langle \frac{\partial V}{\partial x}(t, x(t)) \cdot f(t, x, u) \right\rangle \right\},$$

which is the HJB equation. Notice that $V(x(t), t)$ and $\frac{\partial V}{\partial t}(x(t), t)$ can be moved outside of the infimum, since they are independent of the control $u_{[t,t+\Delta t]}$.

■

Remark. The HJB equation is often solved numerically, since analytic solutions may be difficult or impossible to find.

Appendix A

Appendix to Lecture 12

Below, we present two alternative proofs for the Cayley-Hamilton Theorem. These proofs reveal how the theorem naturally arises from the algebraic and geometric structure of linear operators.

A.1 Cayley-Hamilton Theorem: Alternative Proof 1

For the first alternative proof, we begin by examining Schur's Theorem, which states that, for every finite-dimensional vector space V over an Euclidean space, and linear operator $\mathcal{L} : V \to V$, there exists an orthonormal basis $\mathcal{B}$ of V for which the matrix representation of $\mathcal{L}$ with respect to $\mathcal{B}$ is upper triangular. Below, we present the matrix version.

Theorem A.1 (**Schur's Theorem** ([5], Theorem 6.14, pg. 370))**.** *For any $n \in \mathbb{N}$ and $A \in \mathbb{R}^{n\times n}$, there exists an orthonormal basis $\mathcal{B}$ for $\mathbb{R}^n$ such that $[A]_{\mathcal{B}}$ is upper triangular.*

Proof. The proof follows by induction on n. When $n = 1$, the matrix A is in fact a scalar, so we are done. Suppose the theorem holds for any square matrix in $\mathbb{R}^{k\times k}$, where $k \in \{1, \cdots, n-1\}$. Then the desired result would follow if we can show that there exists an orthonormal basis $\mathcal{B} = \{v_1, \cdots, v_n\}$ for $\mathbb{R}^n$ such that:

$$[A]_{\mathcal{B}} = \begin{bmatrix} B & w_{1:n-1} \\ 0 & w_n \end{bmatrix},$$

for some $B \in \mathbb{R}^{(n-1)\times(n-1)}$, $w_{1:n-1} \equiv \begin{bmatrix} w_1 & \cdots & w_{n-1} \end{bmatrix}^T \in \mathbb{R}^{n-1}$, and $v_n \in \mathbb{R}$. This is because we could then apply the induction hypothesis to $B \in \mathbb{R}^{(n-1)\times(n-1)}$, to generate a suitable orthonormal basis for $\text{span}(\{v_1, \cdots, v_{n-1}\}) = \{v_n\}^{\perp}$), for which the corresponding matrix representation of B is upper triangular, thus completing the proof. (We use the notation $\{v_n\}^{\perp} \equiv \{w | \langle w, v_n\rangle = 0\}$; notice that this is a vector space). This is equivalent to postulating the existence of a vector $v_n \in \mathbb{R}^n$ such that $\{v_n\}^{\perp}$ is A-invariant. (If so, then it makes sense to define the *restriction* of A on $\{v_n\}^{\perp}$. This is, in fact, the matrix B).

We claim that a suitable choice of "v_n," as defined above is any eigenvector of $A^{\star}$ corresponding to any eigenvalue $\lambda \in \sigma(A^{\star})$. In other words, we claim that for such a choice

of v_n, the $(n-1)$-dimensional subspace $\{v_n\}^\perp$ is A-invariant. This is because, for any $x \in \text{span}(\{w\})^\perp$, we have:

$$\langle Ax, w\rangle = \langle x, A^\star w\rangle = \lambda\langle x, w\rangle = 0.$$

The proof is done.

■

Remark. The statement of the theorem given in [5], Theorem 6.14, pg. 370, concerns an arbitrary linear operator L on an arbitrary finite-dimensional vector space $\mathcal{V}$, with no assumptions on the field $\mathcal{F}$ over which $\mathcal{V}$ is defined. In that case, it is unclear whether the characteristic polynomial of L has n roots (where $n \in \mathbb{N}$ is the dimension of $\mathcal{V}$), since the Fundamental Theorem of Algebra only applies to real and complex fields. This is addressed by imposing the additional condition that the characteristic polynomial of L *splits*, i.e. it yields n roots.

Theorem A.2 (**Cayley-Hamilton Theorem**). *Let* $\mathrm{A} \in \mathbb{R}^{n\times n}$*, and suppose its characteristic polynomial has the form:*

$$\chi_{\mathrm{A}}(s) \equiv \det(sI - \mathrm{A}) = s^n + d_1 s^{n-1} + \cdots + d_{n-1}s + d_n$$

Then:

$$\chi_A(A) = A^n + d_1 A^{n-1} + \cdots + d_{n-1}A + d_n I = O$$

Proof. As demonstrated above, Schur's Theorem shows that each square matrix A can be related to at least one upper triangular matrix via a similarity transformation (specifically, via an orthogonal transformation). Since similarity transformations do not change the characteristic equation, we merely have to prove the Cayley-Hamilton Theorem for the case where $A \in \mathbb{R}^{n\times n}$ is upper triangular.

Now, let a_{ij} denote the (i,j)-th element of A. Since A is upper triangular, $a_{ij} = 0$ whenever $i > j$, and:

$$\chi_A(s) = \prod_{k=1}^{n}(s - a_{kk})$$

We wish to show that $\chi_A(A) = O$. This is equivalent to showing that, we have $\chi_A(A)v = 0$ for each $v \in \mathbb{R}^n$, which is in turn follows by showing that:

$$\chi_A(A)\, e_i = \prod_{k=1}^{i}(A - a_{kk}I)e_i = 0, \tag{A.1}$$

for each $i = 1, \cdots, n$, where e_i denotes the i-th standard vector in $\mathbb{R}^n$.

Below, we verify (A.1) via induction on i. When $i = 1$, since A is upper-triangular, we have $(A - a_{11}I)e_1 = 0$, (A.1) holds. Now, suppose for some $i > 1$, (A.1) holds for each

$j = 1, \cdots, i-1$. Again, since A is upper-triangular, we have:

$$\begin{aligned} e_i &= \sum_{j=1}^{i} a_{ji} e_j, \\ (A - a_{ii} I) e_i &= \sum_{j=1}^{i-1} a_{ji} e_j, \\ \Rightarrow \prod_{k=1}^{i} (A - a_{kk} I) e_i &= \prod_{k=1}^{i-1} (A - a_{kk} I) \left(\sum_{j=1}^{i-1} a_{ji} e_j \right) = \sum_{j=1}^{i-1} a_{ji} \left(\prod_{k=1}^{i-1} (A - a_{kk} I) e_j \right) = 0, \end{aligned}$$

since, by applying (A.1) to each $j = 1, \cdots, i-1$, we have $\prod_{k=1}^{j} (A - a_{kk} I) e_j = 0$. Thus, (A.1) holds for i, and by induction, we are done. ■

A.2 Cayley-Hamilton Theorem: Alternative Proof 2

Our second alternative proof to the Cayley-Hamilton Theorem demonstrates that, for a fixed $A \in \mathbb{R}^{n\times n}$, each $v \in \mathbb{R}^n$ can be associated with a particular polynomial factor of $\chi_A(I)$ that annihilates it. Specifically, this polynomial is the characteristic polynomial of A *restricted on the cyclic subspace generated by* v, a concept defined below.

Theorem A.3. *Let $A : \mathbb{R}^n \to \mathbb{R}^n$ be a linear operator, and let $v \in \mathbb{R}^n$ be arbitrarily given. Then $\mathcal{C}_A(v)$, the A-**cyclic subspace generated by** $\mathbf{v}$:*

$$\mathcal{C}_A(v) \equiv \text{span}(\{A^k v \,|\, k = 0, 1, 2, \cdots\})$$

is the smallest A-invariant subspace of $\mathcal{V}$ that contains v.

Note. The span of an infinite set of vectors is defined here as the *finite* linear combination of its elements, to prevent situations in which the resulting subspace is not closed under infinite linear combinations.

Proof. We have four claims to verify:

- $\mathcal{C}_A(v)$ is a subspace of $\mathbb{R}^n$.
- $\mathcal{C}_A(v)$ contains v.
- $\mathcal{C}_A(v)$ is A-invariant.
- $\mathcal{C}_A(v)$ is the smallest subspace of $\mathbb{R}^n$ that satisfies the above two properties, i.e. any other subspace of $\mathbb{R}^n$ that satisfies the above properties contains $\mathcal{C}_A(v)$.

The first two claims follow from the definition of $C_A(v)$ as the span of a collection of vectors that includes v.

To verify the third claim, let $u \in \mathcal{C}_A(v)$ be arbitrarily given. Then there exists $m \in \mathbb{N}$, and scalars $a_0, a_1, \cdots, a_{n-1}$ such that:

$$\begin{aligned} u &= a_0 v + a_1 A v + \cdots + a_m A^m v \\ \Rightarrow\ Au &= a_0 Av + a_1 A^2 v + \cdots + a_{m+1} A^{m+1} v \in \mathcal{C}_A(v), \end{aligned}$$

so $\mathcal{C}_A(v)$ is A-invariant.

To verify the fourth claim, let $\mathcal{W}$ be an arbitrary A-invariant subspace, and let $v \in \mathcal{W}$ be arbitrarily given. We will prove by induction that $\mathcal{C}_A(v) \subset W$. By definition, $A^0 v = v \in \mathcal{W}$. Now, suppose $A^{k-1}v \in \mathcal{W}$ for some $k \in \mathbb{N}$. Since $\mathcal{W}$ is a A-invariant subspace, $A^k v = A(A^{k-1}v) \in \mathcal{W}$. By induction, $A^k v \in \mathcal{W}$ for each $k = 0, 1, 2, \cdots$, so $\mathcal{C}_A(v) \subset \mathcal{W}$. ■

Theorem A.4. *Let $\mathcal{V}$ be a vector space. Using the same terminology as used in the above theorem, we have:*

1. $\dim \mathcal{C}_A(v)$ *is also finite-dimensional, and dim*$\mathcal{C}_A(v) \leq \dim \mathcal{V}$.

2. *Define* $k \equiv \dim \mathcal{C}_A(v) \leq n$. *Then* $\mathcal{B} = \{v, Av, \cdots, A^{k-1}v\}$ *is an ordered basis for* $\mathcal{C}_A(v)$. *Let scalars* $b_0, b_1, \cdots, b^{k-1}$ *be given such that:*

$$A^k v = b_0 v + b_1 Av + \cdots + b_{k-1} A^{k-1} v$$

Then:

$$[A|_{\mathcal{C}_A(v)}]_{\mathcal{B}} = \begin{bmatrix} 0 & 0 & \cdots & 0 & b_0 \\ 1 & 0 & \cdots & 0 & b_1 \\ 0 & 1 & \cdots & 0 & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & b_{k-1} \end{bmatrix} \tag{A.2}$$

In particular,

$$\chi_{A|_{\mathcal{C}_A(v)}}(\lambda) = \lambda^k - b_{k-1}\lambda^{k-1} - \cdots - b_1\lambda - b_0$$

Proof.

1. Since $C_A(v)$ is a subspace of $\mathcal{V}$, it is finite-dimensional, with dimension no greater than $\dim \mathcal{V}$.

2. Let m be the greatest positive integer such that $\mathcal{S} \equiv \{A^{k-1}, \cdots, Av, v\}$ is a linearly independent subset (since $k < \infty$, such an m must exist). Define $\mathcal{W} = \text{span}(\mathcal{S})$; then $\mathcal{S}$ is an ordered basis for $\mathcal{W}$. By definition of m, $\{A^m v\} \cup \mathcal{S}$ is linearly dependent, so $A^m v \in \mathcal{W}$. Then, for each $w \in \mathcal{W}$, there exist scalars $a_0, a_1, \cdots, a_{m-1} \in \mathbb{F}$ such that:

$$\begin{aligned} w &= a_0 v + a_1 Av + \cdots + a_{m-1} A^{m-1} v \\ Lw &= a_0 Av + a_1 A^2 v + \cdots + a_{m-2} A^{m-1} v + a_{m-1} A^m v \end{aligned}$$

Since $v \in \mathcal{W}$, this implies that $\mathcal{W}$ is an A-invariant subspace of $\mathcal{V}$ that contains v. But $\mathcal{C}_A(v)$ is the smallest A-invariant subspace of $\mathcal{V}$ that contains v, so $\mathcal{W} \in \mathcal{C}_A(v)$. On the other hand, $\mathcal{W} \subset \mathcal{C}_A(v)$ by definition, so $\mathcal{W} = \mathcal{C}_A(v)$. This implies that $\mathcal{S}$ is an ordered basis for $\mathcal{C}_A(v)$, so $k = m$, i.e. $\mathcal{S} = \mathcal{B}$. Also:

$$A(A^{k-1}v) = A^k v = b_{k-1} A^{k-1} v + \cdots + b_1 Av + b_0 v$$

and $A(A^j v) = A^{j+1} v$ for each $j = 0, 1, 2, \cdots, k-2$. So (A.2) holds.

It is straightforward to verify that:

$$\chi_{A|_{\mathcal{C}_A(v)}}(\lambda) = \lambda^k - b_{k-1}\lambda^{k-1} - \cdots - b_1\lambda - b_0$$

via induction on (A.2) (a procedure also seen in the derivation of the controllable canonical form).

■

The following lemma, a direct consequence of the Second Representation Theorem (Theorem 4.10), bridges the gap between the theory of cyclic subspaces and the Cayley-Hamilton Theorem.

Lemma A.5. *Let $A : \mathbb{R}^n \to \mathbb{R}^n$ be a linear operator. If $\mathcal{V}_1$ and $\mathcal{V}_2$ are A-invariant subspaces of $\mathbb{R}^n$, and $\mathbb{R}^n = \mathcal{V}_1 \oplus \mathcal{V}_2$, then:*

$$\chi_A(\lambda) = \chi_{A|_{\mathcal{V}_1}}(\lambda) \cdot \chi_{A|_{\mathcal{V}_2}}(\lambda)$$

Proof. Since $\mathcal{V}_1$ and $\mathcal{V}_2$ are A-invariant, and $\mathbb{R}^n = \mathcal{V}_1 \oplus \mathcal{V}_2$ the Second Representation Theorem (Theorem 4.10) tells us that there exist ordered bases $\mathcal{B}_1, \mathcal{B}_2$, and $\mathcal{B}$, of $\mathcal{V}_1, \mathcal{V}_2$, and $\mathbb{R}^n$, respectively, with $\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2$ (in that order), such that:

$$[A]_{\mathcal{B}} = \begin{bmatrix} A_1 & O \\ O & A_2 \end{bmatrix},$$

where $A_1 = A|_{\mathcal{V}_1}$ and $A_2 = A|_{\mathcal{V}_2}$ are the *restrictions* of A on $\mathcal{V}_1$ and $\mathcal{V}_2$, respectively. The lemma follows by observing that:

$$\begin{aligned} \chi_A(\lambda) &= \det(\lambda I - [A]_{\mathcal{B}}) \\ &= \det(\lambda I - [A]_{\mathcal{V}_1}) \cdot \det(\lambda I - [A]_{\mathcal{V}_2}) \\ &= \chi_{A|_{\mathcal{V}_1}}(\lambda) \cdot \chi_{A|_{\mathcal{V}_2}}(\lambda) \end{aligned}$$

■

Theorem A.6 (**Cayley-Hamilton Theorem**). *Let $\mathrm{A} \in \mathbb{R}^{n \times n}$, and suppose its characteristic polynomial has the form:*

$$\chi_{\mathrm{A}}(s) \equiv \det(sI - \mathrm{A}) = s^n + d_1 s^{n-1} + \cdots + d_{n-1} s + d_n$$

Then:

$$\chi_A(A) = A^n + d_1 A^{n-1} + \cdots + d_{n-1} A + d_n I = O$$

Proof. If $n = 0$, then $A = 0$, and the result is evident. Suppose $n \geq 1$, and fix any nonzero $v \in \mathbb{R}^n$. Let $\tilde{A} = A|_{\mathcal{C}_A(v)}$ and $k = \dim \mathcal{C}_A(v) \leq n$. Let scalars $b_0, b_1, \cdots, b_{k-1}$ be given such that $A^k v = b_{k-1} A^{k-1} v + \cdots + b_0 v$, and $[\tilde{A}]_{\mathcal{B}}$ is as expressed in (5.1). We thus have:

$$\begin{aligned} \chi_{\tilde{A}}(\lambda) &= \lambda^k - b_{k-1}\lambda^{k-1} - \cdots - b_1 \lambda - b_0 \\ \Rightarrow \chi_{\tilde{A}}(A)v &= A^k v - b_{k-1} A^{k-1} v - \cdots - b_0 v \end{aligned}$$

Since v is an arbitrary nonzero vector, and $\chi_{\tilde{A}}(A)0 = 0$, then $\chi_{\tilde{A}}(A)v = 0$ for each v $\in \mathbb{R}^n$. By the corollary above, $\chi_{\tilde{A}}(\lambda)$ is a factor of $\chi_A(\lambda)$, $\chi_{\tilde{A}}(\lambda) = g(\lambda)\chi_A(\lambda)$ for some polynomial $g(\lambda)$. We thus have:

$$\chi_A(A)v = g(A)\chi_{\tilde{A}}(A)v = 0,$$

so $\chi_A(A) = O$. ■

Appendix B

Appendix to Lecture 15

Note (**Notation**). In the mathematical statements below, B_r will be used to denote an open ball of radius r centered at the origin of the vector space under consideration (usually $\mathbb{R}^n$).

B.1 Rate of Decay

Proposition B.1 (**Rate of Decay**, [9], Proposition 5.3, pg. 184)**.** *Suppose $x = 0$ is an equilibrium point of the system:*

$$\dot{x} = f(x,t), \qquad x(t_0) = x_0.$$

where f is locally Lipschitz with respect to x in some ball B_h, with Lipschitz constant k, and piecewise continuous with respect to t. Then:

$$|x_0|e^{-k(t-t_0)} \leq |x(t)| \leq |x_0|e^{k(t-t_0)}$$

Proof. Observe that:

$$\begin{aligned}
&\left|\frac{d}{dt}|x|^2\right| = 2|x|\left|\frac{d|x|}{dt}\right|, \\
&\left|\frac{d}{dt}|x|^2\right| = \left|\frac{d}{dt}x^Tx\right| = 2\left|x^T\frac{dx}{dt}\right| \leq 2|x|\left|\frac{dx}{dt}\right|, \\
\Rightarrow &\left|\frac{d|x|}{dt}\right| \leq \left|\frac{dx}{dt}\right|
\end{aligned}$$

Since $f(x,t)$ is Lipschitz continuous, and $f(x,0) = 0$, it follows that:

$$-k|x| \leq \frac{d}{dt}|x| \leq k|x|$$

The desired result follows by applying the Bellman-Gronwell lemma to each of the above inequlities, provided the trajectory stays in the ball B_h in which the Lipschitz condition holds. ■

Remark. The above proposition tells us that a trajectory starting inside B_h will stay inside B_h for a finite amount of time. If $f(x,t)$ is, in fact, globally Lipschitz, it will always stay inside B_h. The proposition also tells us that the rate of convergence of trajectories is at most exponentially.

B.2 Basic Lyapunov Theorems:

Theorem B.2 (**Basic Stability Theorems of Lyapunov**, [9], pgs. 189-192)**.** *The following table associates different notions of internal stability with different conditions on $v(x,t)$ and $\dot{v}(x,t)$. Without loss of generality, the equilibrium point has been placed the origin.*

Table B.1: Basic Lyapunov Theorems

	Conditions on $v(x,t)$	Conditions on $-\dot{v}(x,t)$	Conclusions
1	l.p.d.f.	≥ 0 locally	stable
2	l.p.d.f., decrescent	≥ 0 locally	unif. stable
3	l.p.d.f., decrescent	l.p.d.f.	unif. asymp. stable
4	p.d.f., decrescent	p.d.f.	globally unif. asymp. stable

Proof.

1. Since $v(x,t)$ is locally positive definite and $\dot{v}(x,t) \leq 0$ locally, there exists some $s, r > 0$ and $\alpha(\cdot) \in K$ such that:

$$v(x,t) \geq \alpha(|x|), \qquad \forall x \in B_s, \tag{B.1}$$
$$\dot{v}(x,t) \leq 0, \qquad \forall x \in B_r, \qquad \forall t \geq 0. \tag{B.2}$$

Fix $\epsilon > 0$, and let $\epsilon \equiv \min\{\epsilon, r, s\}$. Since $v(0,t_0) = 0$, and v is continuous (by virtue of it being locally positive definite), there exists some $\delta > 0$ such that:

$$\beta(t_0, \delta) \equiv \sup_{|x|<\delta} v(x,t_0) < \alpha(\epsilon_1) \tag{B.3}$$

Combining (B.1) with (B.3), we have:

$$\alpha(|x(t_0)|) \leq v(x(t_0), t_0) < \alpha(\epsilon_1). \tag{B.4}$$

Since $\alpha(\cdot) \in K$, it follows that $|x(t_0)| < \epsilon_1$.

We now prove that $|x(t_0)| \leq \delta$ implies $|x(t)| < \epsilon_1, \forall t \geq t_0$, which establishes the desired result. Suppose by contradiction that there exists some $t_1 > t_0$ such that $|x(t) \geq \epsilon_1$. (Without loss of generality, we may assume that t_1 is the earliest time satisfying this requirement). Then, from (B.1) and (B.4), we have:

$$v(x(t_0), t_0) \leq \alpha(\epsilon_1) \leq \alpha(|x(t_1)|) \leq v(x(t_1), t_1), \tag{B.5}$$

contradiction the fact that $\dot{v}(x,t) \leq 0$ for all $|x| < \epsilon_1$. Thus:

$$|x(t)| < \epsilon_1, \forall t \geq t_0,$$

establishing the desired claim.

2. Since v is decrescent, the function:

$$\beta(\delta) \equiv \sup_{|x|<\delta} \sup_{t \geq t_0} v(x,t)$$

is non-decreasing; moreover, there exists some $\beta' \in K$ such that, for each $d > 0$:

$$\beta(\delta) \leq \beta'(d), \qquad \forall\, \delta \in (0,d)$$

Now, choose δ such that $\beta(\delta) < \alpha(\epsilon_1)$ (Such a choice can always be made, since β is continuous and $v(0,t_0) = 0$). Applying (B.1), we obtain:

$$\alpha(|x|) \leq v(x,t) \leq \beta(\delta) < \alpha(\epsilon_1), \qquad \forall |x| < \delta, \forall\, t \geq t_0.$$

Since $\alpha(\cdot) \in K$, we have $|x| < \epsilon_1$, as desired.

3. Since $-\dot{v}(x,t)$ is locally positive definite, it satisfies the conditions in the above proof, so 0 is a uniformly stable equilibrium point. We wish to show the existence of some $\delta_1 > 0$ and non-decreasing function $T : \overline{\mathbb{R}^+} \to \overline{\mathbb{R}^+}$, such that for each $\epsilon > 0$, whenever $|x_0| < \delta_1$ and $t > T(\epsilon)$:

$$|\phi(t_1 + t, x_0, t_1)| < \epsilon.$$

(Recall that $\phi(t, x_0, t_0)$ denotes the trajectory of the system $\dot{x} = f(x,t), x(t_0) = t_0$, starting from x_0 at time t_0.)

By hypothesis, $v(x,t)$ is locally positive definite and decrescent, and $-\dot{v}(x,t)$ is locally positive definite, so there exist functions $\alpha(\cdot), \beta(\cdot), \gamma(\cdot)$ such that, whenever $t \geq t_0$ and $|x| < r$:

$$\alpha(|x|) \leq v(x,t) \leq \beta(|x|), \tag{B.6}$$
$$\dot{v}(x,t) \leq -\gamma(|x|). \tag{B.7}$$
$$\tag{B.8}$$

Now, fix $\epsilon > 0$, and define δ_1, δ_2, T such that:

$$\beta(\delta_1) < \alpha(r), \tag{B.9}$$
$$\beta(\delta_2) < \min\{\alpha(\epsilon), \beta(\delta_1)\}, \tag{B.10}$$
$$T = \frac{\alpha(r)}{\gamma(\delta_2)} \tag{B.11}$$

(For a pictoral explanation, see [9], Figure 5.3, pg. 191.)

We now claim that there exists some $t_2 \in [t_1, t_1 + T]$ such that $|\phi(t_2, x_0, t_1)| < \delta_2$. By contradiction, if:

$$|\phi(t, x_0, t_1)| \geq \delta_2, \qquad \forall\, t \in [t_1, t_1 + T], \tag{B.12}$$

we would have:

$$\begin{aligned} 0 \leq \alpha(\delta_2) &\leq v(s(t_1+T, x_0, t_1), t_1+T) \\ &= v(x_0, t_1) + \int_{t_1}^{t_1+T} \dot{v}\big(\phi(\tau, x_0, t_1)\big)\, d\tau \\ &\leq \beta(\delta_1) - T\gamma(\delta_2) \\ &\leq \beta(\delta_1) - T\alpha(r) \\ &< 0, \end{aligned}$$

a contradiction. The first line follows from (B.6) and (B.12), the second from the definition of $\dot{v}(x,t)$ as the time derivative of $v(x,t)$, the third from (B.7) and (B.12), the fourth from (B.11), and the fifth from (B.9).

Now, take t_2 as given in the above claim, and observe that:

$$\begin{aligned} \alpha(|\phi(t, x_0, t_1|) &\leq v(\phi(t, x_0, t_1), t) \leq v(\phi(t_2, x_0, t_1), t_2) \\ &\leq \beta(|\phi(t_2, x_0, t_1)|) \leq \beta(\delta_2) \\ &\leq \alpha(\epsilon), \end{aligned}$$

The first line follows from (B.6), the second from the fact that $t \geq t_1 + T \geq t_2$), the third from (B.6), the fourth from the definition of t_2 as satisfying $|\phi(t_2, x_0, t_1)| < \delta_2$, and the fifth from (B.10).

Since $\alpha(\cdot) \in K$, it follows that $|\phi(t, x_0, t_1)| < \epsilon$ for each $t \geq t_1 + T$. Because $t_1 \geq 0$ was arbitrarily chosen, we are done.

4. Here, we restrict $v(x,t)$ and $-\dot{v}(x,t)$ to be positive definite (not just locally positive definite). As a result, we can reapply the above proof, with $r, \delta \to \infty$, and arrive at the desired conclusion.

■

B.3 Exponential Stability Theorem:

Theorem B.3 (**Exponential Stability Theorem** (Theorem 5.17, pg. 195)). *Suppose* $f(x,t) : \overline{\mathbb{R}^+} \times \mathbb{R}^n \to \mathbb{R}^n$ *has continuous first partial derivatives in* x*, and is piecewise continuous in* t*. Then the following two statements are equivalent:*

1. $x = 0$ *is a locally exponentially stable equilibrium point of* $\dot{x} = f(x,t)$*:, i.e. there exists some* $h, m, \alpha > 0$ *such that for each* $x \in B_h$*:*

$$|\Phi(t, t_0)| \leq m e^{-\alpha(t-t_0)}$$

2. *There exists a function* $v(x,t)$ *and some* $h, \alpha_1, \alpha_2, \alpha_3, \alpha_4 > 0$ *such that:*

$$\alpha_1 |x|^2 \leq v(x,t) \leq \alpha_2 |x|^2 \tag{B.13}$$

$$\frac{dv}{dt}(x,t)\Big|_{\substack{\dot{x}=f(x,t) \\ x(t_0)=x_0}} \leq -\alpha_3 |x|^2 \tag{B.14}$$

$$\left|\frac{\partial v}{\partial x}(x,t)\right| \leq \alpha_4 |x| \tag{B.15}$$

Proof.

"(1) $\Rightarrow$ (2)" : Below, we verify each of the above three inequalities, i.e. (B.13), (B.14), and (B.15), in turn. We start by defining a suitable value function $v(x,t)$ that exploits the exponentially decaying nature of the state:

$$v(x,t) \equiv \int_t^{t+T} |\phi(\tau, x, t)|^2 \, d\tau,$$

where $T > (2/\alpha) \ln m$. (Recall that $\phi(t, x_0, t_0)$ denotes the trajectory of the system $\dot{x} = f(x,t), x(t_0) = t_0$, starting from x_0 at time t_0).

1. (B.13):

   First, observe that since $f(x,t)$ has continuous first partial derivatives with respect to x, the function $f(x,t)$ must be Lipschitz continuous with respect to x. Let k denote the Lipschitz constant. Since the system is exponentially stable with rate α, Theorem B.1 implies there exists some $h > 0$ such that:

$$|x| e^{k(\tau - t)} \leq |\phi(\tau, x, t)| \leq m |x| e^{-\alpha(\tau - t)}$$

   Substituting into the definition of $v(x,t)$ and integrating, we obtain:

$$\frac{1 - e^{-2/T}}{2k} |x|^2 \leq v(x,t) \leq \frac{(1 - e^{-2/T}) m^2}{2\alpha} |x|^2$$

   which establishes (B.13), with:

$$\alpha_1 \equiv \frac{1 - e^{-2/T}}{2k}, \qquad \alpha_2 \equiv \frac{(1 - e^{-2/T}) m^2}{2\alpha}$$

2. (B.14):

Differentiating $v(x,t)$ with respect to t, we have:

$$\begin{aligned}\frac{dv}{dt}(x,t) &= |\phi(t+T,x,t)|^2 - |\phi(t,x,t)|^2 + \int_t^{t+T} \frac{d}{dt}\big(|\phi(\tau,x(t,t)|^2\big)\,d\tau \\ &\leq m^2 e^{-2\alpha T}|x|^2 - |x|^2 + 0 \\ &\leq -(1-m^2e^{-2\alpha T})|x|^2\end{aligned}$$

which, by definition of T, satisfies (B.15), with:

$$\alpha_4 \equiv 1 - m^2 e^{-2\alpha T}$$

3. (B.15):

Differentiating $v(x,t)$ with respect to each component of the state x_i, we have:

$$\frac{\partial v}{\partial x_i}(x,t) = 2\int_t^{t+T} \sum_{j=1}^{n} \phi_j(\tau,x,t)\frac{\partial \phi_j}{\partial x_i}(\tau,x,t)\,d\tau. \tag{B.16}$$

Observe that, since $\phi(\tau,x)$ is smooth and "well-behaved" in general:

$$\begin{aligned}\frac{d}{d\tau}\left(\frac{\partial \phi_i}{\partial x_j}(\tau,x,t)\right) &= \frac{\partial}{\partial x_j}\left(\frac{\partial \phi_i}{\partial \tau}(\tau,x,t)\right) \\ &= \sum_{k=1}^{n} \frac{\partial}{\partial x_k}\left(\frac{\partial \phi_i}{\partial \tau}(\tau,x,t)\right)\cdot \frac{\partial \phi_k}{\partial x_j}(\tau,x,t) \\ &= \sum_{k=1}^{n} \frac{\partial f_i}{\partial x_k}(\tau,x,t)\cdot \frac{\partial \phi_k}{\partial x_j}(\tau,x,t),\end{aligned}$$

along the trajectory satisfying $\dot{x} = f(x,t), x(t_0) = x_0$. If we define:

$$Q_{ij}(\tau,x,t) \equiv \frac{\partial \phi_i}{\partial x_j}(\tau,x,t), \qquad A_{ij}(x,t) \equiv \frac{\partial f_i}{\partial x_j}(x,t)$$

we find that the above equations can be rewritten in the compact form:

$$\frac{d}{d\tau}Q(\tau,x,t) = A(\phi(\tau,x,t),t)\cdot Q(\tau,x,t)$$

In other words, $Q(\tau,x,t)$ is the state transition matrix associated with $A(\phi(\tau,x,t),t)$. Since we assume that the partials of f with respect to x, it follows that there exists some $k > 0$ such that:

$$\begin{aligned}&|A(\cdot,\cdot)| \leq k, \\ \Rightarrow\ &|Q(\tau,x,t)| \leq e^{k(\tau-t)}.\end{aligned}$$

Using this and (B.16), we have:

$$\left|\frac{\partial v}{\partial x}(x,t)\right| \leq 2\int_t^{t+T} m|x|e^{(k-\alpha)(\tau-t)}\,d\tau$$

which satisfies (B.15) with:

$$\alpha_4 \equiv \frac{2m(e^{(k-\alpha)T}-1}{k-\alpha}$$

"(2) $\Rightarrow$ (1)" : This direction is straightforward. From (B.13), (B.14), (B.15), we have:

$$\begin{aligned}
&\dot{v}(x,t) \leq -\frac{\alpha_3}{\alpha_2}v(x,t),\\
\Rightarrow &v(x(t),t) \leq v(x(t_0),t_0)\cdot e^{-\frac{\alpha_3}{\alpha_2}(t-t_0)}\\
\Rightarrow &\alpha_1|x(t)|^2 \leq v(x(t),t) \leq v(x(t_0),t_0)\cdot e^{-\frac{\alpha_3}{\alpha_2}(t-t_0)} \leq \alpha_2|x(t_0)|^2 e^{-\frac{\alpha_3}{\alpha_2}(t-t_0)}\\
\Rightarrow &|x(t) \leq \sqrt{\frac{\alpha_2}{\alpha_1}}|x(t_0)|\cdot e^{-\frac{\alpha_3}{\alpha_2}(t-t_0)}
\end{aligned}$$

■

B.4 Lyapunov Equation: Uniqueness of Solution

Our alternative proof of Lemma 4.48 will involve properties of the *Krnoecker product* of matrices, as derived and explained in detail in Professor Chee-Fai Yung's *Lecture Notes on Mathematical Control Theory* [12]. We begin with its definition below.

Definition B.4 (Kronecker Product). *Given $A = [a_{ij}]_{m\times n} \in \mathbb{C}^{m\times n}$ and $B = [b_{ij}]_{p\times q} \in \mathbb{C}^{p\times q}$, the Kronecker product of A and B is defined as:*

$$A \otimes B \equiv \begin{bmatrix} a_{11}B & a_{11}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}B & a_{n2}B & \cdots & a_{nn}B \end{bmatrix}_{mp\times nq}$$

Example. If $A \in \mathbb{R}^{2\times 2}$ and $B \in \mathbb{R}^{2\times 3}$ are given as:

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, \qquad B = \begin{bmatrix} 2 & 4 & 6 \\ 0 & -1 & 2, \end{bmatrix}$$

then the Kronecker product of A and B is defined as:

$$A \otimes B = \left[\begin{array}{ccc|ccc} 2 & 4 & 6 & 4 & 8 & 12 \\ 0 & -1 & 2 & 0 & -2 & 4 \\ \hline 6 & 12 & 18 & 8 & 16 & 24 \\ 0 & -3 & 6 & 0 & -4 & 7 \end{array}\right] \in \mathbb{R}^{4\times 6}$$

Note. Usually, we denote the i-th row, j-th column element of a matrix $A \in \mathbb{R}^{m\times n}$ by "A_{ij}." For tensor products, we will use a slightly modified version of this standard notation. Given $A \in \mathbb{C}^{m\times n}$ and $B \in \mathbb{C}^{p\times q}$, define:

$$(A \otimes B)_{ii',jj'} \equiv (A \otimes B)_{p(i-1)+i',q(j-1)+j'},$$

where:

$$\begin{aligned} i &= 1, \cdots, m, \\ j &= 1, \cdots, n, \\ i' &= 1, \cdots, p, \\ j' &= 1, \cdots, q, \end{aligned}$$

and the right-hand side of the above expression uses standard notation. In other words, $(A \otimes B)_{ii',jj'}$ denotes the (i', j') in the (i, j) matrix block. This modified notation will be useful in subsequent proofs, where it is convenient to perform matrix multiplications one block matrix at a time. Sometimes, we will mix the two notations:

$$\begin{aligned} (A \otimes B)_{ii',k} &\equiv (A \otimes B)_{i(p-1)+i',k}, \\ (A \otimes B)_{k,jj'} &\equiv (A \otimes B)_{k,j(q-1)+j'} \end{aligned}$$

Again, the right-hand side uses standard notation.

Some basic properties of Kronecker products are presented below.

Proposition B.5 (**Properties of the Kronecker Product**). *For any complex scalars a and complex matrices A, B, C, D:*

1. $(aA) \otimes B = A \otimes (aB) = a(A \otimes B)$.
2. $(A \otimes B)^T = A^T \otimes B^T$.
3. $(A \otimes B)^\star = A^\star \otimes B^\star$.
4. $(A \otimes B) \otimes C = A \otimes (B \otimes C)$.
5. $A \otimes (B + C) = A \otimes B + A \otimes C$.
6. $(B + C) \otimes A = B \otimes A + C \otimes A$.
7. $A \otimes B = O$ *if and only if* $A = O$ *or* $B = O$.
8. *If A, B are both symmetric or both Hermitian, then so is $A \otimes B$.*
9. *If A, B are both upper triangular or both lower triangular, so is $A \otimes B$.*
10. $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$.
11. $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$.

(Clearly, the tenth statement only holds when the dimensions of A, B, C, D are compatible, and the eleventh statement only holds when A, B are invertible).

Proof. Part 1 to Part 9 follow from the definition of the Kronecker product, and in some cases, brute expansion.

For Part 10, observe that:

$$\begin{aligned}
\because \left[(A \otimes B)(C \otimes D\right]_{ii',jj'} &= \sum_{k"} (A \otimes B)_{ii',k"} (C \otimes D)_{k,jj"} \\
&= \sum_{k,k'} a_{ik} b_{i'k'} c_{kj} d_{k'j'} = \left[\sum_k a_{ik} c_{kj}\right] \cdot [b_{i'k'} d_{k'j'}] \\
&= (AC)_{ij} \otimes (BD)_{i'j'} = \left[(AC) \otimes (BD)\right]_{ii',jj'}, \\
\Rightarrow (A \otimes B)(C \otimes D) &= (AC) \otimes (BD)
\end{aligned}$$

For Part 11, we apply Part 10:

$$(A \otimes B)(A^{-1} \otimes B^{-1}) = (AA^{-1}) \otimes (BB^{-1}) = I \otimes I = I$$

■

Next, we extend the definition of a complex polynomial of two variables to the matrix case, using the definition of the Kronecker product.

Definition B.6 (**Polynomial of Two Square Matrices**). *Given a polynomial of $x, y \in \mathbb{C}$:*

$$p(x,y) = \sum_{i,j=0}^{n} p_{ij} x^i \cdot y^j$$

and two square matrices A, B of arbitrary (possibly different) dimensions, define:

$$p(A,B) = \sum_{i,j=0}^{n} p_{ij} A^i \otimes B^j$$

The following theorem describes the relationship between $\sigma(p(A,B))$ and $\sigma(A)$, $\sigma(B)$.

Theorem B.7. *Let $p(x,y)$ be a polynomial of $x, y \in \mathbb{C}$, and let A, B be square matrices of arbitrary (possibly different) dimensions. Then:*

$$\sigma(p(A,B)) = \{p(\lambda_i, \mu_j) | \lambda_i \in \sigma(A), \mu_j \in \sigma(B)\}.$$

Proof. Let A, B be square matrices of arbitrary (possibly different) dimensions. Since $p(A,B)$ is essentially a linear combination of Kronecker products of different powers A and B, we begin by proving the result for $A^i \otimes B^j$, for arbitrary i, j. To that end, let P, Q be invertible matrices such that:

$$J_A \equiv P^{-1} A P,$$
$$J_B \equiv Q^{-1} B Q$$

are the Jordan forms of A, B, respectively. Then, by Part 10 and Part 11 of the above proposition, we have:

$$\begin{aligned} J_A^k \otimes J_B^l &= (P^{-1}AP)^k \otimes (Q^{-1}BQ)^l = (PA^kP^{-1}) \otimes (Q^{-1}B^lQ) \\ &= (P^{-1} \otimes Q^{-1})(A^k \otimes B^l)(P \otimes Q) = (P \otimes Q)^{-1}(A^k \otimes B^l)(P \otimes Q) \end{aligned}$$

Thus, $J_A^i \otimes J_B^j$ and $A^i \otimes B^j$ are similar, and thus share the same eigenvalues. By Part 9 of the above proposition, since J_A and J_B are both upper triangular (by definition of the Jordan form), so is $J_A^i \otimes J_B^j$; its eigenvalues can thus be read off its diagonal, as follows:

$$\sigma(A^i \otimes B^j) = \sigma(J_A^i \otimes J_B^j) = \{\lambda_i^k \mu_j^l | \lambda_i \in \sigma(A), \mu_j \in \sigma(B)\}$$

Now, let $p(x,y) = \sum_{k,l} p_{kl} A^k B^l$ be given. Then:

$$\begin{aligned} p(J_A, J_B) &= \sum_{k,l} p_{kl} J_A^k \otimes J_B^l \\ &= (P \otimes Q)^{-1} \left(\sum_{k,l} p_{kl} A^k \otimes B^l \right) (P \otimes Q) \\ &= (P \otimes Q)^{-1} p(A,B) (P \otimes Q)^{-1} \end{aligned}$$

Thus, $p(J_A, J_B)$ and $p(A, B)$ are similar, so:

$$\sigma(p(A,B)) = \sigma(p(J_A, J_B)) = \{p(\lambda_i, \mu_j) | \lambda_i \in \sigma(A), \mu_j \in \sigma(B)\}$$

■

Corollary B.8. *Let $A \in \mathbb{C}^{n\times n}$ and $B \in \mathbb{C}^{m\times m}$, with:*

$$\sigma(A) = \{\lambda_i | i = 1, \cdots, n\},$$
$$\sigma(B) = \{\mu_j | j = 1, \cdots, m\}.$$

Then, we have:

1. $\sigma(A \otimes B) = \{\lambda_i \mu_j | \lambda_i \in \sigma(A), \mu_j \in \sigma(B)\}$,
2. $\sigma(A \otimes I_m + I_n \otimes B) = \{\lambda_i + \mu_j | \lambda_i \in \sigma(A), \mu_j \in \sigma(B)\}$.

Proof. Both results follow from the above theorem, by taking

1. $p_1(x, y) = xy$, i.e. $p_1(A, B) = A \otimes B$, and
2. $p_2(x, y) = x + y = x^1 y^0 + x^0 y^1$, i.e. $p_2(A, B) = A \otimes I_m + I_n \otimes B$.

■

Remark. Given $A \in \mathbb{C}^{n\times n}$ and $B \in \mathbb{C}^{m\times m}$, the expression $\sigma(A \otimes I_m + I_n \otimes B)$ is sometimes called the *Kronecker sum* of A and B.

Definition B.9 (**Stacking Operator**)**.** *Let $A \in \mathbb{C}^{m\times n}$ be arbitrarily given, and let a_i denote the i-th column of A, i.e.:*

$$A = \begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix}$$

*Define the **stacking operator** vec: $\mathbb{C}^{m\times n} \to \mathbb{C}^{mn}$ by:*

$$\text{vec}(A) = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}.$$

In other words, the stacking operator takes any matrix input and "stacks" up its columns to create a long output vector.

Remark. Observe that vec is linear, i.e. for any $A, B \in \mathbb{C}^{m\times n}$ and $a, b \in$, we have:

$$\text{vec}(aA + bB) = a\text{vec}(A) + b\text{vec}(B)$$

It is also bijective, since it simply "stacks up" the columns of a matrix into a long column vector without changing any of its elements.

The following theorem presents an interesting relationship between the Kronecker product and the vec operator.

Theorem B.10. *Let $A \in \mathbb{C}^{m\times n}$, $B \in \mathbb{C}^{p\times q}$, and $X \in \mathbb{C}^{n\times p}$ be arbitrarily given. Then:*

$$vec(AXB) = (B^T \otimes A)\text{vec}(X)$$

Proof. The proof follows via brute-force expansion. For each $i = 1, \cdots, m$, $j = 1, \cdots, q$:

$$\begin{aligned}
\big(\text{vec}(AXB)\big)_{m(j-1)+i} = (AXB)_{ij} &= \sum_{k=1}^{n}\sum_{k'=1}^{q} A_{ik}X_{kk'}B_{k'j} \\
&= \sum_{k'=1}^{q}\sum_{k=1}^{n} B^T_{jk'}A_{ik}X_{kk'} \\
&= \sum_{k'=1}^{q}\sum_{k=1}^{n} (B^T \otimes A)_{ji,k'k}\big(\text{vec}(X)\big)_{n(k'-1)+k} \\
&= \sum_{l=1}^{np} (B^T \otimes A)_{ji,l}\big(\text{vec}(X)\big)_{l} \\
&= \big((B^T \otimes A)\text{vec}(X)\big)_{m(j-1)+i},
\end{aligned}$$

where we have used the modified notation defined earlier. The desired result follows. ■

Remark. The above theorem allows us to reformulate linear matrix equations. Consider:

$$\sum_{i=1}^{N} A_i X B_i = C,$$

where $A_i \in \mathbb{C}^{m\times n}$, $B_i \in \mathbb{C}^{p\times q}$ for each $i = 1, \cdots, N$, and $X \in \mathbb{C}^{n\times p}$. Applying the vec operator to the above equation and swapping the left- and right-hand sides, we have:

$$\text{vec}(C) = \sum_{i=1}^{N} \text{vec}(A_i X B_i) = \left[\sum_{i=1}^{N} (B_i^T \otimes A_i)\right] \text{vec}(X)$$

which resembles the familiar form of the matrix equation "$Ax = b$." We will use a similar technique below.

Definition B.11 (**Sylvester Equation**)**.** *The Sylvester Equation is of the form:*

$$AX - XB = -C, \tag{B.17}$$

where $A, B, C \in \mathbb{C}^{n\times n}$ are known, while $X \in \mathbb{C}^{n\times n}$ is unknown.

Remark. The Sylvester Equation, which works out to be a system of linear equations, is used in output regulation. It includes the Lyapunov Equation as a special case; indeed, we recover the Lyapunov Equation from the Sylvester Equation by replacing A, B, C with $A^\star, -A^\star, Q$, respectively, then swap A for $A^\star$:

$$A^\star X + XA + Q.$$

The next theorem uses properties of the Kronecker product and stacking operator (specifically, Theorem B.10 and Corollary B.8), to establish *a necessary and sufficient condition under which the Sylvester Equation has a unique solution.*

Theorem B.12. *The Sylvester Equation has a unique solution if and only if A, B share no common eigenvalue, i.e.:*

$$\sigma(A) \cap \sigma(B) = \phi.$$

Proof. Applying the vec operator on both sides of the Sylvester Equation, and applying Theorem B.10, we have the following equivalent (because vec is bijective) equation:

$$(I_n \otimes A - B^T \otimes I_n)\text{vec}(X) = -\text{vec}(C).$$

Thus, the Sylvester Equation has a unique solution if and only if $(I_n \otimes A - B^T \otimes I_n)$ is invertible, i.e. if 0 is not one of its eigenvalues (given by Corollary B.8:

$$\sigma((I_n \otimes A - B^T \otimes I_n) = \{\mu_j - \lambda_i | \lambda_i \in \sigma(A), \mu_j \in \sigma(B)\}.$$

Thus, the following statements are equivalent:

$$\begin{aligned} &\text{The Sylvester Equation has a unique solution} \\ \iff &\sigma((I_n \otimes A - B^T \otimes I_n) \text{ does not contain } 0 \\ \iff &\mu_j \neq \lambda_i, \qquad \forall \lambda_i \in \sigma(A), \mu_j \in \sigma(B) \\ \iff &A, B \text{ have distinct eigenvalues.} \end{aligned}$$

This establishes the theorem. ■

Lemma B.13. *Consider the system $\dot{x} = Ax$, where $A \in \mathbb{R}^{n \times n}$ and $\sigma(A) \in \mathbb{C}^-$. Then the unique solution to Lyapunov's Equation, $A^\star P + PA = -Q$, is given by:*

$$P = \int_0^\infty e^{A^T t} Q e^{At}\, dt \tag{B.18}$$

In particular, the above integral is well-defined.

Proof. We can verify that (4.4) solves the Lyapunov Equation by substituting it into the Lyapunov Equation:

$$\begin{aligned} A^\star X + XA &= \int_0^\infty \left(A^\star e^{A^\star t} Q e^{At} + e^{A^\star t} Q e^{At} A\right) \\ &= \int_0^\infty \frac{d}{dt}\left(e^{A^\star t} Q e^{At}\right)\, dt = e^{A^\star t} Q e^{At}\Big|_0^\infty = -Q, \end{aligned}$$

where the exponential stability of A implies that $\sigma(A) \in \mathbb{C}^-$, so:

$$\lim_{t\to\infty} e^{At} = \lim_{t\to\infty} e^{A^\star t} = O.$$

It remains to show that (4.4) *uniquely* solves the Lyapunov Equation. As remarked above, the Lypaunov Equation is a special case of the Sylvester Equation, as can be seen by taking $B = -A^\star$. Since A is exponentially stable:

$$\begin{aligned} &\begin{cases} \sigma(A) \in \mathbb{C}^-, \\ \sigma(B) = \sigma(-A^\star) = \{-\lambda^\star | \lambda \in \sigma(A)\} \in \mathbb{C}^+, \end{cases} \\ \Rightarrow &\sigma(A) \cap \sigma(B) = \phi. \end{aligned}$$

By Theorem B.12, the solution (4.4) is unique.

■

B.5 Indirect Lyapunov's Method

Theorem B.14 (**Indirect Lyapunov's Method** (Theorems 5.41, 5.42, pgs. 215-217))**.** *Suppose the non-linear system $\dot{x} = f(x,t)$, $x(t_0) = x_0$ has the linear approximation:*

$$\dot{x} = f(x,t) = A(t)x + f_1(x,t), \qquad \text{with} \tag{B.19}$$
$$\lim_{|x|\to 0} \sup_{t\geq 0} \frac{|f_1(x,t)|}{|x|} = 0.$$

Then the following statements hold:

1. *If $\frac{\partial f(x,\cdot)}{\partial x}\big|_{x=0}$ is bounded in time, and 0 is a uniformly asymptotically stable equilibrium point of the linearized system:*

$$\dot{\hat{z}}(t) = \left.\frac{\partial f_1(x,t)}{\partial x}\right|_{x=0} z(t), \tag{B.20}$$

   then 0 is also a locally uniformly asymptotically stable equilibrium point of the original non-linear system $\dot{x} = f(x,t)$, $x(t_0) = x_0$.

2. *If $\frac{\partial f(x,\cdot)}{\partial x}\big|_{x=0}$ is constant in time, and has at least one eigenvalue in $\mathbb{C}^+$, then 0 is an unstable equilibrium point of the original nonlinear system $\dot{x} = f(x,t)$, $x(t_0) = x_0$.*

Proof.

1. Since $A(\cdot) \equiv \frac{\partial f(x,\cdot)}{\partial x}\big|_{x=0}$ is bounded, and 0 is a uniformly asymptotically stable equilibrium point of (B.20), the Time-Varying Lyapunov Lemma (Lemma **??**) states that:

$$P(t) = \int_t^\infty \Phi^\star(\tau,t)\Phi(\tau,t)\,d\tau$$

   is bounded above and below, i.e. $\exists \alpha, \beta > 0$ such that:

$$\alpha|x|^2 \leq x^\star P(t) x \leq \beta |x|^2.$$

   In other words, $v(x,t) \equiv x^\star P(t) x$ is locally positive definite and decrescent. Observe that $P(t)$ is simply as defined in Time-Varying Lyapunov Lemma (Lemma 4.47), except with $Q = I$.

   Meanwhile, by Time-Varying Lemma also implies that $P(t)$ is uniformly bounded in time, i.e. $\sup_{t\geq 0} \|P(t)\| < \infty$. Then (B.19) implies there exists $r > 0$ such that:

$$|f_1(x,t)| \leq \frac{1}{3\sup_{t\geq 0}\|P(t)\|}|x|, \qquad \forall\, x \in B_r,\, \forall\, t \geq 0,$$

Now, we put together the above two facts, and consider the Lie derivative of $v(x,t)$ along $f(x,t)$. or each $x \in B_r$:

$$\begin{aligned}
\dot{v}(x,t) &= \big[Ax + f_1(x,t)\big]^\star P(t)x + x^\star \dot{P}(t)x + x^\star P(t)\big[Ax + f_1(x,t)\big] \\
&= x^\star\big[\dot{P}(t) + P(t)A + A^\star(t)P(t)\big]x + 2x^\star P(t) f_1(x,t) \\
&= -x^\star I x + 2x^\star P(t) f_1(x,t) \\
&\leq -|x|^2 + 2|x| \cdot \|P(t)\| \cdot |f_1(x,t)| \\
&\leq -|x|^2 + \frac{2}{3} \cdot |x|^2 \\
&\leq -\frac{1}{3}|x|^2,
\end{aligned}$$

Thus, $-\dot{v}(x,t)$ is locally positive definite, so 0 is a locally uniformly asymptotically stable equilibrium point of $\dot{x} = f(x,t), x(t_0) = x_0$.

2. From the theorem statement, we know that:

$$A_0 \equiv \left.\frac{\partial f(x,\cdot)}{\partial x}\right|_{x=0}$$

is independent of time. Now, consider the Lyapunov equation:

$$A_0^\star P + P A_0 = I.$$

If $\sigma(A_0) \cap \mathbb{C}^0 = \phi$, then the Taussky Lemma implies that the Lyapunov equation has a unique solution; moreover, since A_0 has at least one eigenvalue in $\mathbb{C}^+$, the unique solution P ahs at least one positive eigenvalue. Then $v(x,t) = x^\star P x$ has positive values arbitrarily close to the origin, and is decrescent (since $v(x,t) \leq \|P\| \cdot |x|^2$).

Now, from (B.19), we know there exists some $r > 0$ such that:

$$|f_1(x,t)| \leq \frac{1}{3\|P\|}|x|, \qquad \forall x \in B_r, \forall t \geq 0,$$

Thus, from the proof of the first part of this theorem:

$$\begin{aligned}
\dot{v}(x,t) &= x^\star\big[PA + A_0^\star P\big]x + 2x^\star P f_1(x,t) \\
&= x^\star I x + 2x^\star P f_1(x,t) \\
&\geq |x|^2 - 2|x| \cdot \|P\| \cdot |f_1(x,t)| \\
&\geq \frac{1}{3}|x|^2.
\end{aligned}$$

Thus, $\dot{v}(x,t)$ is decrescent. By the Basic Instability Theorem (Theorem 4.51), the given system is unstable.

If A_0 has at least one eigenvalue in $\mathbb{C}^-$, and at least another on $\mathbb{C}^0$, the desired result follows by continuity.

■

Bibliography

[1] Apostol, Tom M. *Mathematical Analysis*. Addison-Wesley Publishing Company, Inc., Reading, Massachusetts 2nd Edition, 1974.

[2] Bansal, Somil. *Discussion Notes on Linear Systems Theory*. University of California, Berkeley, Fall 2018.

[3] Bellman, Richard E. *Dynamic Programming*. Princeton University Press, 41 William Street, New Jersey, 08540, 1957.

[4] Desoer, Charles and Callier, Frank. *Linear System Theory*. Springer Science+Business Media, New York, 1991.

[5] Friedberg, Stephen H., Insel, Arnold J., and Spence, Lawrence E. *Linear Algebra*. Pearson Education, Upper Saddle River, New Jersey, 07458, 4th Edition, 2003.

[6] Liberzon, Daniel. *Calculus of Variations and Optimal Control, A Concise Introduction*. Princeton University Press, 41 William Street, New Jersey, 08540, 2012.

[7] Ma, Yi. *Lecture Notes on Linear Systems Theory*. University of Illinois at Urbana-Champaign, Fall 2000.

[8] Rudin, Walter. *Introduction to Mathematical Analysis*. McGraw-Hill, Inc., 3rd Edition, 1976.

[9] Sastry, Shankar. *Nonlinear Systems, Analysis, Stability and Control,* Springer Verlag, 1999.

[10] Tomlin, Claire. *Lecture Notes on Linear Systems Theory*. University of California, Berkeley, Fall 2017.

[11] Yung, Chee-Fai. *Linear Algebra*. Wunan, Taiwan, Second Edition (Chinese Edition), 2012.

[12] Yung, Chee-Fai. *Lecture Notes on Mathematical Control Theory*. National Taiwan University, Taipei, Taiwan, 2013-2014.