

Massimiliano de Sa  
Haeyoon Han

# Lectures on Linear Systems Theory

Draft: February 2025

Caltech Control & Dynamical Systems



# Preface

These lecture notes on linear systems theory are designed for a roughly 10 week course on the subject. The aim is to provide a balanced overview of both the state space and I/O perspectives on linear systems, with an eye towards precise mathematical formulations of problems in control. The approach we follow is inspired by the Fall 2023 offering of CDS 131, Linear Systems Theory, by John Doyle.

Texts that have notably influenced our presentation of the material include *Linear System Theory* by Callier and Desoer, *Mathematical Control Theory* by Sontag, and *Feedback Control Theory* by Doyle, Francis, and Tannenbaum. The lecture notes on linear systems theory by John Lygeros and those by Richard Murray have also proven to be invaluable resources, as has as Stephen Boyd's course EE 363, Linear Dynamical Systems.

Each section of the text corresponds to a roughly 1.5 hour lecture. Subsections that can be skipped without loss of continuity are marked with a ★. To demarcate the difficulty of problems, we also use a star system; ★ means challenging and ★★ means very challenging, compared to the average unstarred problem.

Although the only formal prerequisites for this course are a strong background in linear algebra, prior exposure to control systems and real analysis is certainly useful. Typically, PhD students in control taking this course have either taken or are concurrently taking courses in convex optimization and linear functional analysis, and have had a first (undergraduate) course in control. Prior knowledge of these subjects is not, however, required in order to succeed in learning the course material. We provide a cursory review of the essentials in Chapter 1.

Thanks to the Winter 2025 class of CDS 131—Jedidiah Alindogan, Bhargav Annem, Jennifer Berry, Aditya Bhardwaj, Leo Brenes Calderon, Maven Holst, Mark Hu, Gilchrist Johnson, Minwoo Kim, Dohyun Lee, Xing Hao Li, Derrick Ma, Keyu Wan, Bobby Wang, Meg Wilkinson, Yiheng Xie, Han Xu, Fengqing Yu, Tianyi Zhang, and Catherine Zheng—for crash testing these notes and providing valuable feedback on the course. Special thanks go to Minwoo Kim for close reading!



# Contents

|          |   |           |
|----------|---|-----------|
| <b>1</b> | <b>Mathematical Preliminaries</b>             | <b>7</b>  |
| 1.1      | Vector Spaces                                 | 7         |
| 1.2      | Crumbs of Real Analysis                       | 13        |
| 1.3      | Normed Vector Spaces                          | 16        |
| 1.4      | Banach Spaces                                 | 29        |
| 1.5      | A Refresher on ODEs                           | 33        |
| 1.6      | Further Reading                               | 36        |
| 1.7      | Problems                                      | 36        |
| <b>2</b> | <b>Linear Dynamical Systems</b>               | <b>39</b> |
| 2.1      | Dynamical Systems & State Space Models        | 39        |
| 2.1.1    | Causal Input/Output Dynamical Systems         | 40        |
| 2.1.2    | State Space Representations of Linear Systems | 48        |
| 2.1.3    | Further Reading                               | 52        |
| 2.1.4    | Problems                                      | 53        |
| 2.2      | Solutions of Linear, Time-Varying Systems     | 55        |
| 2.2.1    | Solutions of Continuous-Time Linear Systems   | 55        |
| 2.2.2    | Solutions of Discrete-Time Linear Systems     | 68        |
| 2.2.3    | Further Reading                               | 71        |
| 2.2.4    | Problems                                      | 72        |
| 2.3      | Solutions of Linear, Time-Invariant Systems   | 76        |
| 2.3.1    | Continuous-Time LTI Systems                   | 76        |
| 2.3.2    | Discrete-Time LTI Systems                     | 79        |
| 2.3.3    | The Jordan Canonical Form                     | 80        |
| 2.3.4    | Further Reading                               | 97        |
| 2.3.5    | Problems                                      | 97        |
| 2.4      | Impulse Response & Transfer Functions         | 101       |
| 2.4.1    | Impulse Response of Discrete-Time Systems     | 101       |
| 2.4.2    | Impulse Response of Continuous-Time Systems   | 108       |
| 2.4.3    | Approximations to the Identity ★              | 114       |
| 2.4.4    | The Laplace Transform                         | 118       |
| 2.4.5    | The $\mathcal{Z}$ -Transform                  | 134       |
| 2.4.6    | Further Reading                               | 140       |
| 2.4.7    | Problems                                      | 141       |

|          |   |     |
|----------|---|-----|
| <b>3</b> | <b>Stability of Linear Systems</b>                | 145 |
| 3.1      | The Analytical Approach to Stability              | 145 |
| 3.1.1    | Defining Stability                                | 145 |
| 3.1.2    | Stability of Linear, Time-Varying Systems         | 150 |
| 3.1.3    | Stability of Linear, Time-Invariant Systems       | 153 |
| 3.1.4    | Further Reading                                   | 165 |
| 3.1.5    | Problems  | 165 |
| 3.2      | Invariance-Based Approaches to Stability          | 167 |
| 3.2.1    | The Real Jordan Form                              | 167 |
| 3.2.2    | Phase Portraits                                   | 172 |
| 3.2.3    | The Invariant Subspace Theorem                    | 177 |
| 3.2.4    | The Lyapunov Equation                             | 181 |
| 3.2.5    | Further Reading                                   | 189 |
| 3.2.6    | Problems  | 189 |
| 3.3      | Norms on Signals & Systems                        | 194 |
| 3.3.1    | Poles & Zeros of Transfer Functions               | 194 |
| 3.3.2    | Spaces of Signals                                 | 203 |
| 3.3.3    | Spaces of Systems                                 | 206 |
| 3.3.4    | Further Reading                                   | 217 |
| 3.3.5    | Problems  | 217 |
| <b>4</b> | <b>Fundamental Limits on Linear Systems</b>       | 221 |
| 4.1      | Controllability                                   | 221 |
| 4.1.1    | Defining Controllability & Reachability           | 222 |
| 4.1.2    | Controllability of Linear, Time-Invariant Systems | 227 |
| 4.1.3    | Decompositions & Stabilizability                  | 243 |
| 4.1.4    | Minimum-Energy Control                            | 248 |
| 4.1.5    | Further Reading                                   | 255 |
| 4.1.6    | Problems  | 255 |
|          | <b>References</b>                                 | 259 |

# Chapter 1

## Mathematical Preliminaries

As its name suggests, control *theory* is a fundamentally mathematical subject. At the start of a course in linear systems theory, one is often expected to have mastered topics from linear algebra, real analysis, and functional analysis. Given the modern day engineering curriculum, it's not altogether realistic to expect students to know all of these right from the get-go.

In this brief chapter, we introduce the fundamental mathematical concepts needed to study mathematical systems theory. We stress - you *don't* need to master everything in this chapter upon the first read! If it's your first time seeing this material, try to get a basic feel for the definitions in your first read - you can always come back and refresh your knowledge as the course progresses.

Math is a contact sport, and learning to deal with mathematical abstraction is something that only comes with practice. As such, we leave many of the easier results of this chapter as exercises - you're encouraged to do them as you go along to build your understanding. With this said, let's begin!

### 1.1 Vector Spaces

The fundamental setting for linear system theory is the *vector space*. In this section, we'll get to grips with the basic definitions and properties of vector spaces. Notably, we'll construct very general definitions of vector spaces which don't assume finite-dimensionality.

First, let's construct a formal definition of a vector space. From your first course in linear algebra, you might recall working in vector spaces such as  $\mathbb{R}^n$  or  $\mathbb{C}^n$ , in which *vectors* are tuples of real or complex numbers. You might have defined *vector addition* as the operation which takes two tuples of numbers and adds each pair of entries to form a new vector, or *scalar multiplication*, which multiplies each element of the vector by a scalar to form a new vector. The following definition of a vector space takes the fundamental properties of vectors, addition, and scalar multiplication that are familiar to us from  $\mathbb{R}^n$  and  $\mathbb{C}^n$ , and abstracts away the "tuples of numbers" into an abstract vector.

**Definition 1.1 (Vector Space)** Let  $\mathbb{K} = \mathbb{R}$  or  $\mathbb{C}$ . A vector space  $V$  over  $\mathbb{K}$  is a set,  $V$ , together with two operations,  $+$  :  $V \times V \rightarrow V$  and  $(\cdot)$  :  $\mathbb{K} \times V \rightarrow V$ , satisfying:

1. Closure under operations: For all  $u, v \in V$  and all  $\alpha, \beta \in \mathbb{K}$ ,  $\alpha u + \beta v \in V$ .

2. Associativity: For all  $u, v, w \in V$ ,  $u + (v + w) = (u + v) + w$ .
3. Commutativity: For all  $u, v \in V$ ,  $u + v = v + u$ .
4. Additive Identity: There exists an element  $0 \in V$  for which  $0 + v = v$ , for all  $v \in V$ .
5. Additive Inverse: For all  $v \in V$ , there exists an element  $-v \in V$  for which  $v + (-v) = 0$ .
6. Compatibility: For all  $\alpha, \beta \in \mathbb{K}$  and  $v \in V$ ,  $\alpha \cdot (\beta \cdot v) = (\alpha\beta) \cdot v$ .
7. Multiplicative Identity: For all  $v \in V$ ,  $1 \cdot v = v$ .
8. Distributivity in  $V$ : For all  $\alpha \in \mathbb{K}$  and  $u, v \in V$ ,  $\alpha \cdot (u + v) = \alpha u + \alpha v$ .
9. Distributivity in  $\mathbb{K}$ : For all  $\alpha, \beta \in \mathbb{K}$  and  $v \in V$ ,  $(\alpha + \beta) \cdot v = \alpha v + \beta v$ .

$V$  is called the *set of vectors* and  $\mathbb{K}$  is called the *field*. Elements of the set  $V$  are called *vectors*, while elements of the set  $\mathbb{K}$  are called *scalars*.

*Remark 1.1* Above, we used  $(\cdot)$  to denote the multiplication of a vector by a scalar. One typically suppresses the  $(\cdot)$ , and writes  $c \cdot v = cv$ , for  $c \in \mathbb{K}$  and  $v \in V$ .

*Remark 1.2* Formally, we denote a vector space  $V$  over  $\mathbb{K}$  by the pair  $(V, \mathbb{K})$ . However, if  $\mathbb{K}$  is clear from context, one refers to a vector space simply by the set  $V$ .

Let's take a moment to appreciate what's going on underneath all of the abstraction. First, let's take a moment to outline the structure of an abstract definition, for those who might be unfamiliar with definition-theorem-proof mathematics. Typically, when writing an abstract definition, one starts by specifying out a couple of objects - here, we specify a set,  $V$ , a set  $\mathbb{K}$ , and the operations  $+$  and  $(\cdot)$ . Then, we specify some *axioms* - properties that the objects must have. By laying out each definition in this manner, we can ensure there are no ambiguities in the foundations of our theory.

With this in mind, let's think about what Definition 1.1 is actually saying. All that Definition 1.1 *really* says is that a vector space is any set  $V$  along with a set of scalars  $\mathbb{K}$ , equipped with two operations  $+$  and  $(\cdot)$  that act like  $+$  and  $(\cdot)$  on  $\mathbb{R}^n$  with scalars in  $\mathbb{R}$ . Each condition of the definition simply specifies a property that we have between tuples of numbers in  $\mathbb{R}^n$ , so that our *abstract* vector space behaves just like  $\mathbb{R}^n$  might. The definition is summarized as follows:

1.  $V$  is a set of *vectors*, analogous to tuples of real numbers in  $\mathbb{R}^n$ .
2.  $\mathbb{K}$  is a set of *scalars* called the field, analogous to scalar real numbers in  $\mathbb{R}$ .
3. The operations  $+$  and  $(\cdot)$  are defined to behave just like  $+$  and  $(\cdot)$  between real number scalars and tuples of real numbers in  $\mathbb{R}^n$ .

Let's consider a few examples to make things more concrete.

*Example 1.1* Consider the vector space  $(\mathbb{R}^n, \mathbb{R})$  (read as  $\mathbb{R}^n$  over  $\mathbb{R}$ ). Here,  $V = \mathbb{R}^n$ , the set of tuples of  $n$  real numbers, the set of scalars is  $\mathbb{K} = \mathbb{R}$ , and the operations  $+$  and  $(\cdot)$  are defined,

$$(x_1, \dots, x_n) + (y_1, \dots, y_n) = (x_1 + y_1, \dots, x_n + y_n) \quad (1.1)$$

$$c \cdot (x_1, \dots, x_n) = (c \cdot x_1, \dots, c \cdot x_n), \quad (1.2)$$

where  $c \in \mathbb{R}$  is any scalar in  $\mathbb{R}$  and  $(x_1, \dots, x_n)$ ,  $(y_1, \dots, y_n)$  are tuples of  $n$  real numbers.

*Example 1.2* Consider the vector space  $(\mathbb{C}^{n \times n}, \mathbb{C})$  (read  $\mathbb{C}^{n \times n}$  over  $\mathbb{C}$ ), in which the set of vectors is  $V = \mathbb{C}^{n \times n}$ , the set of  $n \times n$  matrices with complex entries, the set of scalars is  $\mathbb{K} = \mathbb{C}$ , and the operations  $+$  and  $(\cdot)$  are defined,



$$\begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \end{bmatrix} + \begin{bmatrix} b_{11} & \dots & b_{1n} \\ \vdots & \ddots & \vdots \\ b_{n1} & \dots & b_{nn} \end{bmatrix} = \begin{bmatrix} a_{11} + b_{11} & \dots & a_{1n} + b_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} + b_{n1} & \dots & a_{nn} + b_{nn} \end{bmatrix} \quad (1.3)$$

$$c \cdot \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \end{bmatrix} = \begin{bmatrix} c \cdot a_{11} & \dots & c \cdot a_{1n} \\ \vdots & \ddots & \vdots \\ c \cdot a_{n1} & \dots & c \cdot a_{nn} \end{bmatrix}. \quad (1.4)$$

Thus, we observe that the set of  $n \times n$  complex matrices forms a vector space over  $\mathbb{C}$ .

So far, the two examples we've considered have been fairly routine. For the next example, we consider something a little bit more abstract.

*Example 1.3* Let  $V$  be a set and  $W$  a vector space over  $\mathbb{R}$ . Consider the vector space  $(\mathcal{F}_{V,W}, \mathbb{R})$ , where the vectors are functions  $f : V \rightarrow W$ , and the operations  $+$ ,  $(\cdot)$  are defined by function addition and function scalar multiplication,

$$(f + g)(v) = f(v) + g(v), \quad v \in V \quad (1.5)$$

$$(c \cdot f)(v) = c \cdot f(v), \quad v \in V, c \in \mathbb{R}, \quad (1.6)$$

where in the right hand side of the expressions above, the operations  $+$ ,  $(\cdot)$  are those from the vector space  $W$ . Thus, we can generate *new* vector spaces from existing vector spaces! Note that we *don't require*  $V$  to be a vector space for this construction to work, only  $W$ .

This example is certainly more abstract than those we've considered so far, but follows the *same* underlying principles of a vector space. All we have is a set of vectors, a set of scalars, an addition rule, and a scalar multiplication rule. This example highlights a few important principles:

1. Vectors are more than tuples of numbers: Vectors in a vector space are *not* just tuples of numbers - they can take on far more abstract forms such as matrices and functions between vector spaces.
2. Abstraction is your friend: When thinking of vector spaces such as the space of functions above, thinking about the vectors as functions between spaces can get confusing! Instead of *looking inside* the set of vectors, use abstraction to your advantage! When performing any standard algebraic operations, you can *forget* about the complex internal structure, and abstract everything away behind the definition of a vector space. Similarly to computer science, where we hide away implementation details behind classes and methods, we hide away the "implementation" of individual vector spaces (e.g. vectors are *functions*) behind the nice abstraction of Definition 1.1. Using abstraction to your advantage is critical to managing complexity in mathematics.

**Exercise 1.1** Verify that the examples presented above are indeed vector spaces by showing they satisfy all of the axioms of Definition 1.1.

Let's build upon the basic structure of a vector space we outlined above. One of the axioms of a vector space, *closure under operations*, states that for all  $u, v \in V$  and  $\alpha, \beta \in \mathbb{K}$ , one must have  $\alpha u + \beta v \in V$ . It stands to reason that we might like to add or scale more than two vectors at a time! The following definition puts a name to a scaled sum of an arbitrary, finite collection of vectors.

**Definition 1.2 (Linear Combination)** Consider a vector space  $V$  over a field  $\mathbb{K}$ , and a finite collection  $\{v_1, \dots, v_k\} \subseteq V$  of vectors in  $V$ . A linear combination of the collection  $\{v_1, \dots, v_k\}$  is any vector of the form,

$$c_1 v_1 + \dots + c_k v_k, \quad (1.7)$$

where  $c_1, \dots, c_k \in \mathbb{K}$  are any scalars in  $\mathbb{K}$ .

*Remark 1.3* It's important to note that a linear combination of vectors is a scaled sum of a *finite* collection of vectors. We can take as many vectors as we want in a linear combination, as long as the number is not infinite.

There is a subtlety present in the definition of a linear combination that we've glossed over at first pass. The definition of a vector space states that a linear combination of *two* vectors will always belong to the vector space - how do we know that the linear combination of *any* finite linear combination of vectors will be in the vector space? Although this seems like a trivial detail, it's vital that we dot all of our i's and cross all of our t's before we begin using these concepts in earnest. Therefore, every time we make a new definition, we must ensure that the definition *well-posed* - that it is logically sound and doesn't lead to any contradictions in our theory. The following proposition confirms that the definition of a linear combination *is* in fact well-posed.

**Proposition 1.1 (Vector Spaces are Closed Under Linear Combinations)** *Consider a vector space  $V$  and a collection  $\{v_1, \dots, v_k\} \subseteq V$  of vectors in  $V$ . Any linear combination of  $v_1, \dots, v_k$  also belongs to  $V$ .*

**Exercise 1.2** Prove Proposition 1.1. (Hint: use induction on  $k$ ).

Oftentimes, we'll be interested in closely examining a special subset of a given vector space. If the subset of a given vector space *still* has the structure of a vector space, we call it a *subspace*. We make this idea explicit with the following definition.

**Definition 1.3 (Subspace)** Consider a vector space  $V$  over  $\mathbb{K}$  with operations  $+$  and  $(\cdot)$ . A subset  $W \subseteq V$  is said to be a subspace of  $V$  if, under the operations  $+$  and  $(\cdot)$  of  $V$ , it is also a vector space over  $\mathbb{K}$ .

Taking a look at the definition of a subspace, it seems like it would be an awful lot of work to verify that a given subset of a vector space is in fact a subspace. Fortunately, there is a *much* easier way to verify a given set is a subspace than checking that all  $10^n$  axioms of a vector space hold. Since a subspace is already a subset of a vector space, subspaces automatically inherit a number of the vector space properties. What remains to be checked is covered by the following proposition.

**Proposition 1.2** *Consider a vector space  $V$ . A subset  $W \subseteq V$  is a subspace of  $V$  if and only if, for all  $\alpha, \beta \in \mathbb{K}$  and  $u, v \in W$ ,  $\alpha u + \beta v \in W$ .*

In other words, a subset of a vector space is a subspace if and only if it is closed under linear combinations, with respect to the operations of the original vector space. All other properties of a vector space are inherited from the fact that a subspace is a subset of the vector space.

**Exercise 1.3** Prove Proposition 1.2.

An important example of a subspace is the *span* of a collection of vectors, which we now define.

**Definition 1.4 (Span)** Consider a collection  $\{v_1, \dots, v_k\} \subseteq V$  of vectors. The span of  $\{v_1, \dots, v_k\}$ , denoted  $\text{span}\{v_1, \dots, v_k\}$ , is the set of all linear combinations of  $v_1, \dots, v_k$ ,

$$\text{span}\{v_1, \dots, v_k\} = \{c_1 v_1 + \dots + c_k v_k : c_i \in \mathbb{K}\} \subseteq V. \quad (1.8)$$

**Exercise 1.4** Show that the span of a finite collection of vectors of a vector space  $V$  is a subspace of  $V$ .

Let's study some more basic properties of vector spaces and linear combinations. As you may recall, an important property that a collection of vectors can have is *linear independence*. Fundamentally, a collection of vectors is linearly independent if each vector “points in a new direction” compared to the other vectors in the collection. We formalize this intuitive notion of linear independence in the following definition.

**Definition 1.5 (Linear Independence)** Let  $V$  be a vector space over  $\mathbb{K}$ . Consider a collection  $\{v_1, \dots, v_k\} \subseteq V$  of vectors in  $V$ . The collection is said to be *linearly independent* if, for all  $c_1, \dots, c_k \in \mathbb{K}$  with  $c_1, \dots, c_k$  not all zero,  $c_1 v_1 + \dots + c_k v_k \neq 0$ . If the collection is *not* linearly independent, it is said to be *linearly dependent*.

Let's verify that the formal definition of linear independence matches up with our intuitive notion of vectors “pointing in new directions.” Suppose we're given a collection of linearly *dependent* vectors,  $\{v_1, \dots, v_k\}$ . Then, by Definition 1.5, there exist constants  $c_1, \dots, c_k \in \mathbb{K}$ , not all zero, for which

$$c_1 v_1 + c_2 v_2 + \dots + c_k v_k = 0. \quad (1.9)$$

Without loss of generality, suppose that  $c_1 \neq 0$  (the choice to focus on  $c_1$  is arbitrary, as no  $c_i, v_i$  pair has any property making it more “special” than the others - we can therefore focus on  $c_1$  *without loss of generality*). Then, by the above, we can write,

$$v_1 = -\frac{c_2}{c_1} v_2 - \dots - \frac{c_k}{c_1} v_k. \quad (1.10)$$

Therefore, we conclude that  $v_1$  “points in a direction” that is *already* covered by the remaining  $k - 1$  vectors. Thus, it is *not* true that in a linearly dependent set, every vector “points in a new direction.” We conclude that the formal definition of linear independence is consistent with the intuitive notion we outlined above.

Armed with the definition of linear independence, we now have the ability to study a variety of more sophisticated constructions. First, we introduce the definition of a basis for a vector space.

**Definition 1.6 (Basis)** Consider a vector space  $V$  over  $\mathbb{K}$ . A basis  $\mathcal{B}$  for  $V$  is a linearly independent collection,  $\mathcal{B} = \{v_1, \dots, v_k\} \subseteq V$ , such that for all  $v \in V$ , there exist scalars  $c_1, \dots, c_k \in \mathbb{K}$  for which

$$v = c_1 v_1 + \dots + c_k v_k. \quad (1.11)$$

Any vector  $v_i \in \mathcal{B}$  belonging to the basis is called a basis vector.

**Exercise 1.5** Let  $\mathcal{B} = \{v_1, \dots, v_k\}$  be a basis for a vector space  $V$ . Show that for any  $v \in V$ , the constants  $c_1, \dots, c_k \in \mathbb{K}$  for which  $v = c_1v_1 + \dots + c_kv_k$  are uniquely determined by  $v$ .

Thus, a collection of vectors is a *basis* for a vector space if the collection is linearly independent and we can write any element of the vector space as a (unique) linear combination of the basis vectors. Eagle-eyed readers will note that here, we've defined a basis to be a *finite* collection of vectors! We'll address our reasoning for this shortly.

In order to understand why we've defined a basis in terms of a finite collection of vectors, we first need to define the *dimension* of a vector space. Recall that, given a set  $S$  with a finite number of elements,  $|S|$  denotes the number of elements in  $S$ . With this in mind, we make the following definition.

**Definition 1.7 (Dimension)** Consider a vector space  $V$  over  $\mathbb{K}$ . Suppose  $\mathcal{B} \subseteq V$  is a basis for  $V$ . The dimension of  $V$ ,  $\dim(V)$ , is the number of elements in the basis,  $\dim(V) = |\mathcal{B}|$ .

In order for this definition to be well-defined, it's important that we check that *all bases* for a given vector space have the same dimension! The following result confirms that this is in fact the case.

**Proposition 1.3 (Dimension is Well-Defined)** Consider a vector space  $V$  over  $\mathbb{K}$ . If  $\mathcal{B}_1$  and  $\mathcal{B}_2$  are two bases for  $V$ , then  $|\mathcal{B}_1| = |\mathcal{B}_2|$ .

Let's discuss why Proposition 1.3 implies dimension is a well-posed quantity. Given any two bases for  $V$ , Proposition 1.3 states that the bases must contain the same number of elements. Since dimension is defined as the number of elements in a basis, Proposition 1.3 confirms that dimension is a property of a vector space, rather than a property of a basis. Thus, it makes sense to write  $\dim(V)$ , rather than  $\dim(\mathcal{B})$ .

**Exercise 1.6** Prove Proposition 1.3.

Before we move on to the study of linear transformations, we have one more point to discuss regarding bases! In your first course in linear algebra, you likely worked exclusively with *finite-dimensional* vector spaces - vector spaces in which there exists a finite basis. What, then, does it mean for a vector space to be infinite-dimensional?

**Definition 1.8 (Finite/Infinite-dimensional Vector Space)** A vector space  $V$  is said to be finite-dimensional if it has a basis with a finite number of elements. If no such basis exists,  $V$  is said to be infinite-dimensional.

Defining bases for infinite-dimensional spaces is a somewhat more subtle problem; one that is beyond the scope of our brief review of linear algebra. We refer the interested reader to the references at the end of the chapter for a treatment of infinite-dimensional bases.

Now that we've studied vector spaces in reasonable detail, we can discuss linear transformations, which are special maps between vector spaces. From your first course in linear algebra, you might immediately think of a matrix when you think of a linear transformation. If the vector space you're working in is  $\mathbb{R}^n$ , this is certainly justifiable! However, in abstract vector spaces, matrices are *not* immediately associated with linear transformations, analogous to how vectors are not immediately associated with tuples of numbers. Consider the following, abstract definition.

**Definition 1.9 (Linear Transformation)** Consider two vector spaces  $V, W$  over  $\mathbb{K}$ . A linear transformation between  $V$  and  $W$  is a map  $A : V \rightarrow W$  such that for all  $\alpha, \beta \in \mathbb{K}$  and  $u, v \in V$ ,  $A(\alpha u + \beta v) = \alpha A(u) + \beta A(v)$ .

*Remark 1.4* Note that, instead of writing  $A(u)$  for the action of a linear transformation  $A : V \rightarrow W$  on a vector  $u \in V$ , it is convention to write  $Au$ . This convention *does not* extend to nonlinear maps.

Thus, an “abstract” linear transformation is simply a map between vector spaces that *respects* the linear structure of the vector spaces. The study of linear transformations between vector spaces is one that is surprisingly deep. We’ll return to linear transformations after a brief digression into *analysis* on the real line and in vector spaces.

## 1.2 Crumbs of Real Analysis

In systems and control theory, bounds are used to estimate complex quantities in terms of simple ones. For instance, in Chapter 2, when we study stability, we’ll look for *exponential bounds* on the trajectories of our system. To effectively study mathematical control theory, it’s therefore important that we have the tools we need to bound sets of real numbers, and reason about when these bounds are *sharp*.

This is where real analysis comes in. Fundamentally, elementary real analysis is the centered around the study of *convergence* of sequences and the *shapes* of sets of real numbers. In this section, we’ll focus on this second point, and develop the necessary concepts required to find sharp bounds on sets of real numbers.

First, we’ll introduce some basic language and facts about subsets of the real line,  $\mathbb{R}$ , and then proceed to develop sharp bounds - called suprema and infima - on these sets. Finally, we’ll briefly look at how these sharp bounds on sets interact with functions. As a first step towards achieving these goals, we define what it means for a set to be *bounded*.

**Definition 1.10 (Bounded Above/Below)** Consider a subset  $A \subseteq \mathbb{R}$ .  $A$  is said to be:

1. Bounded above: if there exists an  $U \in \mathbb{R}$  such that  $x \leq U$ , for all  $x \in A$ . In this case, such a  $U$  is said to be an upper bound on  $A$ .
2. Bounded below: if there exists an  $L \in \mathbb{R}$  such that  $L \leq x$ , for all  $x \in A$ . In this case, such an  $L$  is said to be a lower bound on  $A$ .
3. Bounded: if there exists an  $M \in \mathbb{R}$  such that  $|x| \leq M$ , for all  $x \in A$ .

The definitions presented above simply tell us whether or not a set is bounded. Notably, we do *not* specify how close or far our upper or lower bounds are from the set. For instance, if one takes the interval  $[0, 1) \subseteq \mathbb{R}$ , both 1 and 100 are equally valid upper bounds. However, the upper bound of 1 *clearly* provides us with more information about the set than the upper bound of 100. Now, we seek a bound on a set that provides us with the *most possible information* about a set - i.e. a bound that is as tight as possible. Consider the following definitions.

**Definition 1.11 (Supremum)** Consider a set  $A \subseteq \mathbb{R}$ . The supremum of  $A$ , denoted  $\sup A$ , is the least upper bound of  $A$ . That is, if  $U \in \mathbb{R}$  is any upper bound on  $A$ ,  $\sup A \leq U$ .

Just as we define the least upper bound, we also define the greatest lower bound.

**Definition 1.12 (Infimum)** Consider a set  $A \subseteq \mathbb{R}$ . The infimum of  $A$ , denoted  $\inf A$ , is the greatest lower bound of  $A$ . That is, if  $L$  is any lower bound on  $A$ , then  $L \leq \inf A$ .

The supremum and infimum are defined to be the *tightest possible* upper and lower bounds on a given subset of  $\mathbb{R}$ . Let's look at a couple of quick examples.

*Example 1.4* Consider the set  $A = [a, b) \subseteq \mathbb{R}$ , where  $a < b$ . Here,  $\sup A = b$ , as  $b$  is the smallest possible upper bound on  $A$ . Likewise,  $\inf A = a$ , since  $a$  is the greatest possible lower bound on  $A$ .

This example highlights an important feature of suprema and infima - it is *not* necessarily the case that  $\sup A \in A$  or  $\inf A \in A$ . In the example of  $[a, b)$ , the infimum belonged to the set, while the supremum did not. Let's take a second look at the definitions of the supremum and infimum to see why this is. When looking at the definitions of suprema and infima, one makes a natural comparison to the maximum and minimum of a set. Let's make a formal definition of a maximum and minimum to clear up the difference between a maximum/supremum and minimum/infimum.

**Definition 1.13 (Maximum/Minimum)** Consider a subset  $A \subseteq \mathbb{R}$ . A point  $a \in A$  is said to be the maximum element of  $A$  if  $x \leq a$  for all  $x \in A$ . Likewise, a point  $b \in A$  is said to be the minimum element of  $A$  if  $b \leq x$  for all  $x \in A$ .

Thus, we observe that unlike the supremum and infimum, the maximum or minimum of a set *must belong* to the set. As a consequence of this, given an arbitrary subset  $A \subseteq \mathbb{R}$ , one is *not* guaranteed to have a maximum or minimum value, even if it is bounded above and below! Consider, for example, the set  $(a, b) \subseteq \mathbb{R}$ . Here, the set is bounded above and below *but* has no maximum or minimum value, since  $a$  and  $b$  are not included. However, both the supremum and infimum exist;  $\sup(a, b) = b$  and  $\inf(a, b) = a$ .

In the event where they exist, how do the maximum and minimum of a set relate to the supremum and infimum? The following result provides an answer.

**Proposition 1.4** Consider a set  $A \subseteq \mathbb{R}$ . If  $A$  has a maximum element, then  $\max A = \sup A$ . Likewise, if  $A$  has a minimum element, then  $\min A = \inf A$ .

Thus, as our intuition might confirm, the maximum and minimum coincide with the supremum and infimum *when they exist*.

**Exercise 1.7** Prove Proposition 1.4.

The question of existence of maxima and minima naturally begs the question - do the supremum and infimum of a set always exist? The following fact answers this question.

**Fact (Axiom of the Supremum)** A nonempty, bounded above set  $A \subseteq \mathbb{R}$  has a finite supremum,  $\sup A < \infty$ .  $\square$

Notably, we state this result as a *fact*, not as a proposition! Why is this? As it happens, the axiom of the supremum is baked into the formal definition of the real numbers,  $\mathbb{R}$ .<sup>1</sup> As such, it is not something that we formally need to prove. In the case where  $A$  is *not* bounded above, we take  $\sup A = \infty$  by convention. In the case where  $A = \emptyset$ , we take  $\sup A = -\infty$  by convention.

This answers the question of existence of the supremum. What about existence of the infimum? In order to answer this question, we state a handy proposition which lets us translate results about the supremum to results about the infimum.

<sup>1</sup> This is a component of one of a number of equivalent, formal definitions for  $\mathbb{R}$ . The interested reader is encouraged to consult the references provided at the end of the chapter to learn more about the construction of the reals.

**Proposition 1.5** For a set  $A \subseteq \mathbb{R}$ ,  $\inf A = -\sup(-A)$ , where  $-A := \{-x : x \in A\}$ .

This result enables us to directly translate results about the supremum to results about the infimum, simply by flipping the sign of elements in the set. Generally, we'll prove results for the supremum and translate them to the infimum using Proposition 1.5.

**Exercise 1.8** Prove Proposition 1.5.

By Proposition 1.5 and the Axiom of the Supremum, the following result is immediate.

**Proposition 1.6** Any nonempty, bounded below set  $A \subseteq \mathbb{R}$  has a finite infimum.

Using our conventions for the supremum along with Proposition 1.5, we have that  $\inf A = -\infty$  when  $A$  is not bounded below and that  $\inf A = \infty$  when  $A = \emptyset$ .

Thus far, we've only discussed the suprema and infima of generic subsets of  $\mathbb{R}$ . Now, we'll consider how suprema and infima interact with real-valued functions. Although this might initially seem like a step up from working with suprema and infima of sets, all we need to do to treat suprema and infima of functions is introduce a little bit of notation. Consider an arbitrary set  $A$  (not even necessarily a subset of  $\mathbb{R}$ ), and a function  $f : A \rightarrow \mathbb{R}$ . We define,

$$\sup_{x \in A} f(x) := \sup\{f(x) : x \in A\} = \sup f(A) \quad (1.12)$$

$$\inf_{x \in A} f(x) := \inf\{f(x) : x \in A\} = \inf f(A). \quad (1.13)$$

Thus, taking the supremum and infimum of real-valued functions is really the same thing as taking the supremum and infimum of sets - we simply take the supremum and infimum of the *images* of sets, which are nothing more than standard subsets of  $\mathbb{R}$ . Now that we've discussed the foundational aspects of suprema and infima, we state a number of their basic properties.

**Proposition 1.7 (Properties of Suprema)** Consider sets  $A, B \subseteq \mathbb{R}$ . The suprema of  $A$  and  $B$  satisfy the following properties.

1. *Subset Inequality:* If  $B \subseteq A$ , then  $\sup B \leq \sup A$ .
2. *Sum of Sets Equality:* If  $A, B$  are nonempty, then  $\sup(A + B) = \sup A + \sup B$ , where  $A + B$  is defined,  $A + B = \{a + b : a \in A, b \in B\}$ .
3. *Sum of Functions Inequality:* For  $D$  an arbitrary set and  $f, g : D \rightarrow \mathbb{R}$  functions,

$$\sup_{x \in D} (f(x) + g(x)) \leq \sup_{x \in D} f(x) + \sup_{x \in D} g(x). \quad (1.14)$$

It's extremely important that we distinguish between properties (2) and (3) listed above. Although it might initially seem that the same rule would apply for functions and sets, this is not the case! The intuitive reasoning for this is as follows. For functions, taking the sum  $\sup_{x \in D} f(x) + g(x)$  means taking the supremum of the sum where  $f$  and  $g$  are evaluated at the *same* point  $x$ . However, taking the sum  $\sup_{x \in D} f(x) + \sup_{x \in D} g(x)$  means that  $f$  and  $g$  can take in different values! This yields an inequality, as both  $f$  and  $g$  are free to individually take on their suprema. Note that this reasoning takes a little bit of sharpening up to yield a formal proof - in particular, one can formally prove (3) by applying (1) and (2). We leave the details of this task as an exercise.

**Exercise 1.9** Apply Proposition 1.5 to reformulate Proposition 1.7 in terms of infima.

**Exercise 1.10** Supply a formal proof of item (3) of Proposition 1.7 using items (1) and (2).

### 1.3 Normed Vector Spaces

Now that we've sketched out some basic real analysis, we return to the domain of linear algebra and study basic analysis in vector spaces. First, we define a norm on a vector space.

**Definition 1.14 (Norm)** Consider a vector space  $V$  over  $\mathbb{K}$ . A norm is a map  $\|\cdot\| : V \rightarrow \mathbb{R}_{\geq 0}$  satisfying the following conditions:

1. Positive Definite:  $\|u\| \geq 0$  for all  $u \in V$ , and  $\|u\| = 0$  if and only if  $u = 0$ .
2. Positive Homogeneity: For all  $u \in V$  and  $c \in \mathbb{K}$ ,  $\|cu\| = |c| \|u\|$ .
3. Triangle Inequality: For all  $u, v \in V$ ,  $\|u + v\| \leq \|u\| + \|v\|$ .

*Remark 1.5* Here, we use  $|\cdot|$  to denote the magnitude of a scalar. For  $\mathbb{K} = \mathbb{R}$ , this is equal to the absolute value, and for  $\mathbb{K} = \mathbb{C}$ , this is equal to the complex magnitude.

Since  $\|\cdot\| : V \rightarrow \mathbb{R}_{\geq 0}$  maps to the positive reals (which *do not* include  $\infty$ ), it is a requirement that the norm of any given vector is finite! If  $\|v\|$  is not finite for some  $v \in V$ , then  $\|\cdot\|$  is *not* a valid norm on  $V$ . Let's consider a few common examples of norms on  $\mathbb{R}^n$ . In each of the following examples, convince yourself that each norm is finite for all vectors in the vector space on which they are defined.

*Example 1.5 ( $\ell^2$  Norm)* The  $\ell^2$  norm on  $\mathbb{R}^n$ , alternatively called the 2-norm or Euclidean norm, is defined

$$\|x\|_2 = \sqrt{x_1^2 + \dots + x_n^2} = \sqrt{\sum_{i=1}^n x_i^2}. \quad (1.15)$$

*Example 1.6 ( $\ell^1$  Norm)* The  $\ell^1$  norm on  $\mathbb{R}^n$ , alternatively called the 1-norm, is defined

$$\|x\|_1 = |x_1| + \dots + |x_n| = \sum_{i=1}^n |x_i|. \quad (1.16)$$

*Example 1.7 ( $\ell^\infty$  Norm)* The  $\ell^\infty$  norm on  $\mathbb{R}^n$ , alternatively called the  $\infty$ -norm, is defined

$$\|x\|_\infty = \max_{i \in \{1, \dots, n\}} |x_i|. \quad (1.17)$$

**Exercise 1.11** Show that the  $\ell^1, \ell^2, \ell^\infty$  norms proposed above are indeed norms on  $\mathbb{R}^n$ .

With the definition of a norm in hand, we're ready to define a *normed vector space*. Scary as this might sound, the definition of a normed vector space is actually quite innocent.

**Definition 1.15 (Normed Vector Space)** A normed vector space is a pair  $(V, \|\cdot\|)$  of a vector space  $V$  and a norm  $\|\cdot\|$  on  $V$ .

That was easy! Let's consider a couple of basic examples of normed vector spaces.

*Example 1.8*  $(\mathbb{R}, |\cdot|)$ , the real line equipped with the absolute value, is perhaps the simplest example of a normed vector space. Similarly,  $(\mathbb{C}, |\cdot|)$ , the complex numbers equipped with the complex magnitude, also forms a normed vector space. Note that if we refer to  $\mathbb{R}$  or  $\mathbb{C}$  as normed vector spaces we will *always* assume the norms in question are the absolute value and complex magnitude, respectively, unless directed otherwise.



*Example 1.9*  $(\mathbb{R}^n, \|\cdot\|_1)$ ,  $(\mathbb{R}^n, \|\cdot\|_2)$ , and  $(\mathbb{R}^n, \|\cdot\|_\infty)$  are all normed vector spaces.

*Example 1.10*  $(\mathbb{R}^{n \times n}, \|\cdot\|_2)$ , the vector space of  $n \times n$  matrices with real values, along with the matrix 2-norm,  $\|A\|_2 = \sigma_{\max}(A)$  is a normed vector space.

*Example 1.11*  $(\mathbb{R}^{n \times n}, \|\cdot\|_F)$ , the vector space of  $n \times n$  matrices with real values, along with the Frobenius norm,  $\|A\|_F = \sqrt{\text{tr}(A^\top A)}$ , is a normed vector space.

These examples are all fairly simple in nature - we take a vector space like  $\mathbb{R}^n$  or  $\mathbb{R}^{n \times n}$ , which we understand well, and place a norm on it that is easy to compute in terms of the entries of the vector. Shortly, we'll discuss some more complex, infinite-dimensional examples. Before we can do this, however, we first need to learn some more analysis.

When we first took a crack at real analysis, we studied different subsets of the real line. We'll begin our study of analysis in normed vector spaces in the same spirit. We start by introducing a special subset of a normed vector space, called an  $\epsilon$ -ball.

**Definition 1.16 (Epsilon Ball)** Consider a normed vector space  $(V, \|\cdot\|)$ . Let  $\epsilon > 0$  be a fixed real number and  $v \in V$  a vector. The  $\epsilon$ -ball centered at  $v$  is the set,

$$B_\epsilon(v) := \{u \in V : \|u - v\| < \epsilon\}. \quad (1.18)$$

*Remark 1.6* An  $\epsilon$ -ball is also commonly referred to as an  $\epsilon$ -neighborhood.

Let's make a few comments on this definition. First, we note that an epsilon ball *doesn't* include points that exactly a distance  $\epsilon$  away from  $v$ ; rather it only includes points that are strictly *less* than  $\epsilon$  away from  $v$ . Thus, an  $\epsilon$  ball has a “fuzzy” boundary. Secondly, we note that the actual shape of an  $\epsilon$ -ball depends on the choice of norm! For instance, an  $\epsilon$ -ball in  $(\mathbb{R}^2, \|\cdot\|_2)$  will look like an actual ball, while an  $\epsilon$  ball in  $(\mathbb{R}^2, \|\cdot\|_\infty)$  will look like a square.

**Exercise 1.12** Sketch the epsilon ball  $B_1(0)$  on a pair of coordinate axes in the spaces  $(\mathbb{R}^2, \|\cdot\|_1)$ ,  $(\mathbb{R}^2, \|\cdot\|_2)$ , and  $(\mathbb{R}^2, \|\cdot\|_\infty)$ . *Hint: the boundaries of these sets should be “fuzzy.”*

Above, we mentioned that an  $\epsilon$  ball has a “fuzzy” boundary - i.e.  $B_\epsilon(v)$  does not have a sharp boundary where the set ends. We know that  $\epsilon$ -balls are not the only sets with “fuzzy” boundaries we can draw - any set without a sharp boundary fits into this category. How, then, can we specify a general class of sets in a normed vector space with a “fuzzy” boundary? Consider the following definition.

**Definition 1.17 (Open Set)** Let  $(V, \|\cdot\|)$  be a normed vector space. A subset  $A \subseteq V$  is said to be an open set if, for all  $v \in V$ , there exists an  $\epsilon > 0$  such that  $B_\epsilon(v) \subseteq A$ .

*Remark 1.7* Note that a few expressions are commonly used to declare a set  $A \subseteq V$  is open. The expressions “ $A$  is an open set,” “ $A$  is open” (if the space  $V$  is clear), or “ $A$  is open in  $V$ ” (if one wishes to emphasize the vector space), are all equivalently used to declare that a set  $A$  is open.

Thus, we declare a set  $A \subseteq V$  to be *open* if, around each point in  $A$ , we can squeeze in an  $\epsilon$ -ball that is still contained in the set. Open sets are the natural way of making precise the idea of a set with a “fuzzy” boundary. Let's run through a couple of examples of open sets.

*Example 1.12 (Examples of Open Sets)*

1. Any open interval,  $(a, b)$ ,  $a < b$ , is an open set in  $(\mathbb{R}, |\cdot|)$ .

2. Any union of open intervals,  $(a, b) \cup (c, d)$ ,  $a < b$ ,  $c < d$ , is an open set in  $(\mathbb{R}, |\cdot|)$ .
3. In a normed vector space  $(V, \|\cdot\|)$  any open ball  $B_\epsilon(v)$  is open.

**Exercise 1.13** Confirm that the examples above are indeed open sets. Sketch out each example and confirm that the sets have “fuzzy boundaries.”

Now that we’ve seen some basic examples of open sets, let’s outline some basic properties of open sets.

**Proposition 1.8 (Properties of Open Sets)** *Let  $(V, \|\cdot\|)$  be a normed vector space. The open subsets of  $V$  satisfy the following properties:*

1. *Nothing & everything:*  $\emptyset$  and  $V$  are open sets.
2. *Stability under unions:* For  $\{U_\alpha\}_{\alpha \in A}$  an arbitrary collection of open sets, the union  $\bigcup_{\alpha \in A} U_\alpha$  is an open set.
3. *Stability under finite intersections:* For  $\{U_i\}_{i=1}^k$  a finite collection of open sets, the intersection  $\bigcap_{i=1}^k U_i$  is an open set.

*Remark 1.8* Just like we write  $\{U_i\}_{i=1}^k$  to refer to the collection  $\{U_1, \dots, U_k\}$  of  $k$  sets, we use the notation  $\{U_\alpha\}_{\alpha \in A}$  to refer to an arbitrary collection of sets, indexed by an arbitrary set  $A$ . Here,  $\alpha$  is the index of an individual set in the collection (analogous to  $i$ ), and  $A$  is the set of all indices of the collection (analogous to  $\{1, \dots, k\}$ ). For instance, if one has a collection of sets corresponding to real numbers, one might write  $\{U_\alpha\}_{\alpha \in \mathbb{R}}$ . When the index set is clear or unimportant, we will simply write  $\{U_\alpha\}$  as shorthand to refer to an arbitrary collection of sets.

Let’s review what each condition of the proposition says. The first condition, *nothing & everything*<sup>2</sup>, states that the empty set (nothing) and the entire vector space  $V$  (everything) are both open sets. The empty set trivially satisfies Definition 1.17, since it has no points to check for, and  $V$  satisfies Definition 1.17 since any  $\epsilon$ -ball is automatically contained in  $V$ .

The next condition, stability under unions, states that an *arbitrary* (potentially uncountably infinite) collection of open sets has a union that is also open. Why is this? If we pick a point  $v \in \bigcup_\alpha U_\alpha$ , the definition of a union tells us there exists an  $\alpha$  for which  $v \in U_\alpha$ . Since  $U_\alpha$  is open, there exists a ball  $B_\epsilon(v) \subseteq U_\alpha \subseteq \bigcup_\alpha U_\alpha$ . Thus, the union is open.

The final condition, stability under finite intersections, states that if we have a *finite* collection of open sets, their intersection must be open. The formal argument for this case takes a little more thought than stability under unions - we leave its proof as an exercise (with a hint) below. Why can’t we take infinite intersections? Consider the following counterexample. Define a collection  $\{B_{1/n}(0)\}_{n \in \mathbb{N}}$ , of  $\epsilon$ -balls centered at the origin with shrinking radius  $1/n$ . With a little work, one may show that,

$$\bigcap_{n=1}^{\infty} B_{1/n}(0) = \{0\}, \quad (1.19)$$

since all of the sets in the collection shrink down towards zero as  $n \rightarrow \infty$ . Since  $\{0\}$  isn’t an open set (we can’t fit an epsilon ball of positive radius around 0 into the set  $\{0\}$ ), this yields an example of an infinite collection of open sets whose intersection is *not* open. This results in condition (3) holding only for finite intersections.

<sup>2</sup> The nice terminology “nothing and everything” is due to Joel Tropp.

**Exercise 1.14** Prove item (2) of Proposition 1.8. Hint: to pick the radius of an  $\epsilon$  ball that fits in the intersection, try experimenting with the *minimum* of a set of epsilons.

Now that we've defined an open set, a natural question to ask is - what, if anything, is a *closed* set? If an open set is a set with a fuzzy boundary, perhaps a closed set should be a set with a sharp boundary. Although this intuitive definition covers a variety of closed sets, it isn't quite expressive enough to capture everything we need (for instance, are  $\emptyset$  and  $V$  closed sets?). Consider the following, abstract definition.

**Definition 1.18 (Closed Set)** Let  $(V, \|\cdot\|)$  be a normed vector space. A subset  $A \subseteq V$  is said to be a closed set if its complement,  $A^c = V \setminus A$ , is an open set.

Let's think for a moment about why this definition might match up with our intuition regarding closed sets having "sharp" boundaries. If a set has a fuzzy boundary, then it stands to reason that its complement should have a sharp boundary. Likewise, if a set has a sharp boundary, then its complement should have a fuzzy boundary. Thus, a set with a sharp boundary should be closed.

**Exercise 1.15** Draw some pictures to reconcile the "sharp boundary" intuition behind closed sets with the formal definition.

Let's explore some basic properties of closed sets.

**Proposition 1.9 (Examples & Properties of Closed Sets)** Let  $(V, \|\cdot\|)$  be a normed vector space. The closed subsets of  $V$  satisfy the following:

1. *Nothing & Everything:*  $\emptyset$  and  $V$  are closed sets.
2. *Stability under intersections:* For  $\{C_\alpha\}_{\alpha \in A}$  an arbitrary collection of closed sets, the intersection  $\cap_{\alpha \in A} C_\alpha$  is a closed set.
3. *Stability under finite unions:* For  $\{C_i\}_{i=1}^k$  a finite collection of closed sets, the union  $\cup_{i=1}^k C_i$  is a closed set.

Thus, we observe that closed sets seem to satisfy the exact *opposite* properties of open sets! The reason for this is precisely that closed sets are defined as complements of open sets - the complement *flips* the properties of open sets to properties of closed sets.

**Exercise 1.16** Use DeMorgan's laws to prove Proposition 1.9 directly from Proposition 1.8.

**Exercise 1.17** Produce an infinite collection of closed sets whose union is not closed.

Closed sets have a number of convenient properties that makes them easy to work with when considering matters such as continuity and convergence. As such, given an arbitrary subset  $A$  of a normed vector space  $V$ , we often want to find the "smallest" closed set containing  $A$ . This way, we can preserve the basic structure of  $A$  while gaining the extra properties of a closed set. Consider the following definition, which defines the "smallest" closed set containing any given set.

**Definition 1.19 (Closure)** Let  $(V, \|\cdot\|)$  be a normed vector space and  $A \subseteq V$  an arbitrary subset. The closure of  $A$ , denoted  $\overline{A}$ , is the smallest closed set containing  $A$ ,

$$\overline{A} = \bigcap_{\alpha \in A} C_\alpha, \text{ where } \{C_\alpha\}_{\alpha \in A} = \{C_\alpha \subseteq V : A \subseteq C_\alpha \text{ and } C_\alpha \text{ is closed}\}. \quad (1.20)$$

As the closure of  $A$  is defined as the *intersection* of all closed sets containing  $A$ , one can think of the closure of  $A$  as “shrink wrapping” the set  $A$  with closed sets. The following two exercises provide some quick sanity checks regarding the closure.

**Exercise 1.18** Verify that the closure of any given set is in fact closed.

**Exercise 1.19** Show that a set  $A \subseteq V$  is closed if and only if  $A = \overline{A}$ .

The following is a nice consequence of these two exercises.

*Example 1.13* Consider a normed vector space  $(V, \|\cdot\|)$ . The closure of any open ball  $B_\epsilon(x)$  in  $V$  is the closed ball,

$$\overline{B}_\epsilon(x) = \{y \in V : \|x - y\| \leq \epsilon\}. \quad (1.21)$$

Thus, the closure makes the ordinarily “fuzzy” boundary of an open ball a “sharp” boundary. Try sketching out the intersections of some closed sets containing the ball  $B_\epsilon(x)$  to convince yourself that this also follows from the definition of the closure.

Just as we can define the smallest closed set containing a given set, we can define the *largest open set* contained within a given set.

**Definition 1.20 (Interior)** Let  $(V, \|\cdot\|)$  be a normed vector space and  $A \subseteq V$  an arbitrary subset. The interior of  $A$ , denoted  $A^\circ$ , is the largest open set contained in  $A$ ,

$$A^\circ = \bigcup_{\alpha \in A} O_\alpha, \text{ where } \{O_\alpha\}_{\alpha \in A} = \{O_\alpha \subseteq V : O_\alpha \subseteq A \text{ and } O_\alpha \text{ is open}\}. \quad (1.22)$$

Instead of “shrink wrapping” a set with a collection of closed sets, as we did in the case of the closure, we can think of the interior as inflating a collection of open sets inside the given set, until we fill up the entire inside space with open sets. What remains after this operation is the largest open set contained inside the given set.

**Exercise 1.20** Verify that the interior of any given set is in fact open.

**Exercise 1.21** Show that a set  $A \subseteq V$  is open if and only if  $A = A^\circ$ .

*Example 1.14* Consider a normed vector space  $(V, \|\cdot\|)$ . The interior of any closed ball  $\overline{B}_\epsilon(x)$  in  $V$  is the open ball,  $B_\epsilon(x)$ . Thus, the interior makes the ordinarily “sharp” boundary of a closed ball a “fuzzy” boundary.

So far, we’ve only examined normed vector spaces  $(V, \|\cdot\|)$  with a fixed choice of norm on  $V$ . Now, we examine what happens when we change the norm on a given vector space. What effect does the choice of norm  $\|\cdot\|$  on  $V$  have on the properties of the normed vector space  $(V, \|\cdot\|)$ ? As we saw in our definition of an open set above, the choice of a norm on a vector space *entirely* determines which sets are open and closed. Thus, to determine the effect of a norm on the vector space, it’s logical to ask - when do two norms on a single vector space determine the same open sets?

As a first step towards answering this question, we define what it means for two norms on the same vector space to be *equivalent*.

**Definition 1.21 (Equivalent Norms)** Consider a vector space  $V$ . Two norms,  $\|\cdot\|_a$  and  $\|\cdot\|_b$ , on  $V$  are said to be equivalent if there exist constants  $k_1, k_2 > 0$  for which

$$k_1 \|v\|_a \leq \|v\|_b \leq k_2 \|v\|_a, \quad \forall v \in V. \quad (1.23)$$

Thus, two norms on a vector space are said to be equivalent if one norm can be “sandwiched” between some positive multiples of the other. We now show that the equivalence of two norms *entirely* determines whether the open sets of the normed vector spaces determined by the norms will be the same.

**Proposition 1.10 (Equivalent Norms Determine the Same Open Sets)** *Consider a vector space  $V$  and norms  $\|\cdot\|_a$  and  $\|\cdot\|_b$  on  $V$ . The open sets of  $(V, \|\cdot\|_a)$  and  $(V, \|\cdot\|_b)$  are the same if and only if  $\|\cdot\|_a$  and  $\|\cdot\|_b$  are equivalent norms.*

**Proof** First, suppose  $\|\cdot\|_a$  and  $\|\cdot\|_b$  are equivalent. Consider an arbitrary open ball  $B_{\epsilon_0}^a(x)$ , with respect to the norm  $\|\cdot\|_a$ . By the definition of norm equivalence, it follows that there exists an open ball  $B_{\epsilon_1}^b(x)$  with respect to the norm  $\|\cdot\|_b$ , satisfying  $B_{\epsilon_1}^b(x) \subseteq B_{\epsilon_0}^a(x)$ .

Now, let  $U_a$  be an arbitrary open set of  $(V, \|\cdot\|_a)$ . We aim to show that  $U_a$  is also open in  $(V, \|\cdot\|_b)$ . Let  $x \in U_a$  be arbitrary. Since  $U_a$  is open in  $x$ , there exists a ball  $B_{\epsilon}^a(x) \subseteq U_a$ . But, by the reasoning above, there exists a ball  $B_{\epsilon}^b(x) \subseteq B_{\epsilon}^a(x) \subseteq U_a$ . So, around every point in  $U_a$ , we can squeeze in a ball defined with respect to  $\|\cdot\|_b$ . We conclude that  $U_a$  is open in  $(V, \|\cdot\|_b)$ . To show that any open set  $U_b$  in  $(V, \|\cdot\|_b)$  is also open in  $(V, \|\cdot\|_a)$ , one follows the same reasoning.

Now, we prove the opposite direction. For this direction, all we have to work with are open sets. We’ll have to be a little bit clever, and pick our open sets such that they directly give us the constants we need for norm equivalence. Suppose the open sets of  $(V, \|\cdot\|_a)$  and  $(V, \|\cdot\|_b)$  are the same - i.e. a set  $U \subseteq V$  is open in  $(V, \|\cdot\|_a)$  if and only if it is open in  $(V, \|\cdot\|_b)$ . Consider the open ball of radius 1 around the origin in  $\|\cdot\|_b$ , denoted  $B_1^b(0)$ . Since  $B_1^b(0)$  is open in  $\|\cdot\|_b$ , it must also be open in  $\|\cdot\|_a$ . Therefore, there exists a ball  $B_{\epsilon_1}^a(0)$  in  $\|\cdot\|_a$  satisfying  $B_{\epsilon_1}^a(0) \subseteq B_1^b(0)$ . There also exists a ball  $B_{\epsilon_2}^a(0)$  in  $\|\cdot\|_a$  satisfying  $B_1^b(0) \subseteq B_{\epsilon_2}^a(0)$ . The inclusion of balls,

$$B_{\epsilon_1}^a(0) \subseteq B_1^b(0) \subseteq B_{\epsilon_2}^a(0), \quad (1.24)$$

implies that for any vector  $v \in V$  satisfying  $\|v\|_b = 1$ ,

$$\epsilon_1 \|v\|_a \leq 1 \leq \epsilon_2 \|v\|_a. \quad (1.25)$$

Now, consider an arbitrary, nonzero vector  $v \in V$ . By the positive homogeneity property of norms, it follows that,

$$\epsilon_1 \left\| \frac{v}{\|v\|_b} \right\|_a \leq \frac{\|v\|_b}{\|v\|_b} \leq \epsilon_2 \left\| \frac{v}{\|v\|_b} \right\|_a \quad (1.26)$$

$$\frac{\epsilon_1}{\|v\|_b} \|v\|_a \leq \frac{\|v\|_b}{\|v\|_b} \leq \frac{\epsilon_2}{\|v\|_b} \|v\|_a \quad (1.27)$$

$$\epsilon_1 \|v\|_a \leq \|v\|_b \leq \epsilon_2 \|v\|_a. \quad (1.28)$$

For the remaining case of  $v = 0$ , the final inequality above holds trivially. We conclude that  $\|\cdot\|_a$  and  $\|\cdot\|_b$  are in fact equivalent.  $\square$

The next result - the proof of which we leave to the problems at the end of the section - is one of the most fundamental results in analysis in normed vector spaces. It tells us that, in any finite-dimensional vector space, *all* norms are equivalent. Because of this fact, the choice of norm in a finite-dimensional vector space often isn’t critical - any two norms on

a finite-dimensional vector space will produce normed vector spaces with similar analytical properties.

**Theorem 1.1 (Norm Equivalence in Finite Dimensions)** *Consider a finite-dimensional vector space  $V$ . Any two norms  $\|\cdot\|_a$  and  $\|\cdot\|_b$  on  $V$  are equivalent.*

**Proof** See [4] for the details.  $\square$

Now that we've discussed the structure of some important subsets of normed vector spaces, we turn our attention to the *continuity* of maps between normed vector spaces. Consider the following definition.

**Definition 1.22 (Continuity)** Consider two normed vector spaces,  $(V, \|\cdot\|_V)$  and  $(W, \|\cdot\|_W)$ . A mapping  $f : V \rightarrow W$  is continuous at a point  $x \in V$  if, for all  $\epsilon > 0$ , there exists a  $\delta > 0$  (possibly a function of  $\epsilon$  and  $x$ ) for which

$$\|x - y\|_V < \delta \implies \|f(x) - f(y)\|_W < \epsilon. \quad (1.29)$$

If  $f : V \rightarrow W$  is continuous at all  $x$  in a subset  $U \subseteq V$ ,  $f$  is said to be continuous on  $U$ . Likewise, if  $f$  is continuous at all  $x \in V$ , it is simply said to be continuous.

Let's briefly check our understanding of this definition. Definition 1.22 states that, if a mapping  $f : V \rightarrow W$  is continuous,  $f(x)$  can change by a small amount if  $x$  changes by a small amount. It's important to note - the norm on  $x - y$  is the norm  $\|\cdot\|_V$  (since  $x, y \in V$ ), while the norm on  $f(x) - f(y)$  is the norm  $\|\cdot\|_W$  (since  $f(x), f(y) \in W$ ). One must take special care to ensure the correct norms are being used on  $x - y$  and  $f(x) - f(y)$  - these norms will *not* in general be the same!

**Proposition 1.11 (Properties of Continuous Functions)** *Let  $(U, \|\cdot\|_U)$ ,  $(V, \|\cdot\|_V)$ , and  $(W, \|\cdot\|_W)$  be normed vector spaces over  $\mathbb{K}$  and  $f : V \rightarrow W$ ,  $g : V \rightarrow W$ , and  $h : W \rightarrow U$  be continuous mappings.*

1. *Algebraic Combinations:* For any continuous functions  $\alpha, \beta : V \rightarrow \mathbb{K}$ , the function  $p : V \rightarrow W$ , defined  $p(x) = \alpha(x) \cdot f(x) + \beta(x) \cdot g(x)$ , is also continuous.
2. *Composition:* The composition  $h \circ f : V \rightarrow U$  is continuous.

**Remark 1.9** In item (1) of the proposition above, we use continuous functions  $\alpha, \beta : V \rightarrow \mathbb{K}$ . Here, we treat  $\mathbb{K}$  as the normed vector space  $(\mathbb{K}, \|\cdot\|)$ , where  $\|\cdot\|$  represents either absolute value (for  $\mathbb{K} = \mathbb{R}$ ) or complex magnitude (for  $\mathbb{K} = \mathbb{C}$ ). It's necessary to treat  $\mathbb{K}$  as a normed vector space in order to apply Definition 1.22!

When studying a function  $f : V \rightarrow \mathbb{R}$  from a normed vector space  $V$  to the real line  $\mathbb{R}$ , it's useful to have some criteria to determine whether or not the function attains a maximum or minimum value on a given set. A special class of sets - termed *compact sets* - enable exactly this ability, among many others. In order to provide a sufficiently abstract definition of a compact set, we first require the definition of an open cover.

**Definition 1.23 (Cover/Open Cover)** Consider a normed vector space  $(V, \|\cdot\|)$  and a subset  $A \subseteq V$ . A collection  $\{U_\alpha\}_{\alpha \in A}$  of subsets of  $V$  is said to be a *cover* of  $A$  if

$$A \subseteq \bigcup_{\alpha \in A} U_\alpha. \quad (1.30)$$

If each  $U_\alpha$  is open, the collection  $\{U_\alpha\}_{\alpha \in A}$  is said to be an *open cover* of  $A$ . A cover is said to be *finite* if it contains a finite number of sets.

*Remark 1.10* Sometimes, one will encounter the phrase, “ $\{U_\alpha\}_{\alpha \in A}$  covers  $A$ .” This is simply another way of saying that  $\{U_\alpha\}_{\alpha \in A}$  is a cover of  $A$ .

This definition tells us that a *cover* of a given set is a collection of sets that, when pasted together, contain the given set.

**Definition 1.24 (Subcover)** Consider a normed vector space  $(V, \|\cdot\|)$ , a subset  $A \subseteq V$ , and a cover  $\{U_\alpha\}_{\alpha \in A}$  of  $A$ . A subcollection  $\{V_\beta\}_{\beta \in B} \subseteq \{U_\alpha\}_{\alpha \in A}$  is said to be a subcover of  $\{U_\alpha\}_{\alpha \in A}$  if  $\{V_\beta\}_{\beta \in B}$  is still a cover of  $A$ . A subcover is said to be *finite* if it contains a finite number of sets.

*Remark 1.11* It’s extremely important to note that  $\{V_\beta\}_{\beta \in B} \subseteq \{U_\alpha\}_{\alpha \in A}$  *does not* mean that the sets  $V_\beta$  are subsets of the sets  $U_\alpha$ . Here, the subset relation is a relation on the *collections*  $\{V_\beta\}$  and  $\{U_\alpha\}$  - we pick out a few of the elements of the collection  $\{U_\alpha\}$  to form  $\{V_\beta\}$ . If the indices of the sets are not changed when one picks these elements out, one will have  $B \subseteq A$ .

*Remark 1.12* Notice how the use of the prefix *sub* mirrors that of a subspace. Generally a sub-“object” is a subset of an “object” that retains the object’s key properties. This is true of a subspace, wherein we have a subset of vector space that remains a vector space, and of a subcover, wherein we have a subset of a cover that remains a cover.

Thus, a subcover of a cover of a given set takes picks out a few sets from the cover that, when pasted together, still cover the given set. Armed with these definitions, we’re ready to state an abstract definition of a compact set.

**Definition 1.25 (Compact Set)** Let  $(V, \|\cdot\|)$  be a normed vector space. A subset  $K \subseteq V$  is said to be a compact set if every open cover of  $K$  has a finite subcover.

*Remark 1.13* To say that a set  $K \subseteq V$  is a compact set, one will often say “ $K$  is compact,” or “ $K$  is compact in  $V$ .” This is just like how, to declare a set  $O$  is an open set, we said “ $O$  is open,” or “ $O$  is open in  $V$ .”

At first glance, this definition seems entirely mystifying. In order to pull back the curtain on compactness, we’ll study a few basic results of compact sets. The following result tells us a few properties that compact sets must satisfy in a normed vector space. Note that the proofs of the next few compactness results are generally out of scope of our treatment of the material - one may consult the references at the end of the chapter for their proofs.

**Proposition 1.12 (Compact Sets are Closed & Bounded)** Let  $(V, \|\cdot\|)$  be a normed vector space and  $K \subseteq V$  be a compact subset of  $V$ . Then,  $K$  must satisfy the following two properties:

1. Closed:  $K$  is closed in  $V$ .
2. Bounded:  $K$  is bounded in  $V$  - there exists an  $M \geq 0$  such that  $\|x\| \leq M$  for all  $x \in K$ .

It’s natural to wonder whether the converse of the result above is true - is every closed & bounded set compact? In *finite-dimensional* normed vector spaces, this turns out to be true, but in infinite-dimensional spaces, this is generally false. We postpone a more complete characterization of the infinite-dimensional case to our later discussion of Banach spaces. For now, we content ourselves with the (very nice) finite-dimensional result.

**Proposition 1.13 (Compact Sets in Finite-Dimensional Spaces)** *Consider a normed vector space  $(V, \|\cdot\|)$ . If  $V$  is finite-dimensional, then a subset  $K \subseteq V$  is compact if and only if it is closed & bounded.*

This proposition produces a quick and easy way to verify a given set in a finite-dimensional vector space is compact. In the following two examples, we illustrate how easy it is to determine some simple subsets of a finite-dimensional space are compact.

*Example 1.15* In  $\mathbb{R}$ , any closed interval  $[a, b]$ ,  $a \leq b$  is compact, since any closed interval is closed and bounded.

*Example 1.16* In  $\mathbb{R}^n$ , any closed box  $[a, b]^n = \{x \in \mathbb{R}^n : a \leq x_i \leq b\}$ ,  $a \leq b$ , is compact, since closed boxes are closed and bounded.

*Example 1.17* In a finite-dimensional normed vector space  $(V, \|\cdot\|)$ , any closed ball  $\overline{B}_\epsilon(x) = \{y \in V : \|x - y\| \leq \epsilon\}$  is compact, since closed balls are closed and bounded.

As alluded to earlier, one of the most appealing features of compact sets is that they interact well with continuous functions. In particular, a continuous function *always* achieves a maximum and minimum value on a compact set. In the following proposition, we summarize some of the most useful properties.

**Proposition 1.14 (Compactness & Continuity)** *Consider two normed vector spaces,  $(V, \|\cdot\|_V)$  and  $(W, \|\cdot\|_W)$ .*

1. *Let  $f : V \rightarrow W$  be a continuous map. For any compact set  $K \subseteq V$ , the image of  $K$  under  $f$ ,  $f(K) \subseteq W$ , is also compact.*
2. *Let  $f : V \rightarrow \mathbb{R}$  be a continuous map. For any compact set  $K \subseteq V$ , the set  $f(K)$  has a maximum and a minimum value. Consequently,*

$$\sup_{x \in K} f(x) = \max_{x \in K} f(x) \text{ and } \inf_{x \in K} f(x) = \min_{x \in K} f(x). \quad (1.31)$$

Property (2) is particularly useful when bounding the values of a function over a given set. Oftentimes, one can find a compact set containing the given set, and use the maxima and minima on the compact set to bound the function values on the given set. The fact that a continuous function achieves a maximum and a minimum value over a compact set means that there is *no risk* of the function “blowing up” to  $+\infty$  or  $-\infty$  on the set.

The following exercise presents a nice consequence of this proposition.

**Exercise 1.22** Let  $(V, \|\cdot\|_V)$  and  $(W, \|\cdot\|_W)$  be normed vector spaces and  $f : V \rightarrow W$  be a continuous map. Show that on any compact set  $K \subseteq V$ , the function  $g : V \rightarrow \mathbb{R}$ , defined  $g(x) = \|f(x)\|_W$ , attains a maximum and a minimum value. *Hint: are norms continuous functions?*

Definition 1.22 offers one definition of continuity of a map between normed vector spaces. What else might we be interested in? Oftentimes, the form of continuity presented in Definition 1.22 is not strong enough! Next, we consider a stronger form of continuity, called Lipschitz continuity.

**Definition 1.26 (Lipschitz Continuity)** Let  $(V, \|\cdot\|_V)$  and  $(W, \|\cdot\|_W)$  be two normed vector spaces. A mapping  $f : V \rightarrow W$  is said to be Lipschitz continuous on  $V$  if there exists a constant  $L \geq 0$ , called a Lipschitz constant, for which

$$\|f(x) - f(y)\|_W \leq L \|x - y\|_V, \quad \forall x, y \in V. \quad (1.32)$$



*Remark 1.14* Unlike continuity, we specify Lipschitz continuity as a global property of a function. Lipschitz continuity is defined over the entire space, rather than at a single point.

By nature of the name Lipschitz *continuity*, it's natural to expect a Lipschitz continuous function to be continuous in the sense of Definition 1.22. In the next proposition, we confirm that our expectations are met.

**Proposition 1.15** *Consider a function  $f : V \rightarrow W$  between normed vector spaces  $(V, \|\cdot\|_V)$  and  $(W, \|\cdot\|_W)$ . If  $f$  is Lipschitz continuous on  $V$ , then it is continuous on  $V$ .*

**Proof** Suppose  $f : V \rightarrow W$  is Lipschitz continuous. We must show that  $f$  is continuous at every  $x \in V$ . First, fix a point  $x \in V$  and a number  $\epsilon > 0$ . To show that  $f$  is continuous at  $x$ , we must find a number  $\delta > 0$  - possibly dependent on  $\epsilon$  and  $x$ , such that  $\|x - y\|_V < \delta$  implies  $\|f(x) - f(y)\|_W < \epsilon$ . In order to identify such a  $\delta$ , we appeal to the definition of Lipschitz continuity. By Definition 1.26, there exists a constant  $L \geq 0$  such that for all  $x, y \in V$ ,  $\|f(x) - f(y)\|_W \leq L \|x - y\|_V$ . Looking at the inequalities,

$$\|f(x) - f(y)\|_W \leq L \|x - y\|_V \quad (\text{what we have}) \quad (1.33)$$

$$\|f(x) - f(y)\|_W < \epsilon \quad (\text{what we want}), \quad (1.34)$$

it seems reasonable that choosing  $\delta = \epsilon/L$  will meet our needs. Let's check that this value works. For any  $y$  satisfying  $\|x - y\|_W < \delta = \epsilon/L$ , we have

$$\|f(x) - f(y)\|_W \leq L \|x - y\|_V < L \cdot \frac{\epsilon}{L} = \epsilon. \quad (1.35)$$

Thus, we conclude that  $f$  is continuous at  $x$ . Since  $x$  was chosen arbitrarily, we conclude that  $f$  is continuous on  $V$ .  $\square$

Above, we showed that all Lipschitz continuous functions are continuous. What about the other direction? Are all continuous functions Lipschitz? As the following exercise illustrates, Lipschitz continuity is a *strictly stronger* condition than continuity.

**Exercise 1.23** The function  $f : \mathbb{R} \rightarrow \mathbb{R}$ , defined  $f(x) = x^2$  is well-known to be continuous. Show that it is *not* Lipschitz continuous. *Hint: proceed by contradiction.*

An important example of a Lipschitz function on any normed vector space  $(V, \|\cdot\|)$  is the norm  $\|\cdot\|$  itself, as a function from  $V \rightarrow \mathbb{R}$ .

**Proposition 1.16** *For  $(V, \|\cdot\|)$  a normed vector space,  $\|\cdot\| : V \rightarrow \mathbb{R}$  is Lipschitz continuous.*

This result implies that the norm of a normed vector space is a continuous function.

**Exercise 1.24** Prove Proposition 1.23.

So far in our discussion of continuity, we've dealt only with *arbitrary* mappings between normed vector spaces. What can we say about the continuity of *linear* transformations? Consider, for example, the simple linear transformation from  $\mathbb{R} \rightarrow \mathbb{R}$ , defined  $f(x) = ax$ ,  $a \in \mathbb{R}$ . It's clear that this function is continuous; in fact, it is Lipschitz continuous with Lipschitz constant  $|a|$ . Therefore, as an initial guess, it seems not too far-fetched to suggest that linear transformations between normed vector spaces are Lipschitz continuous.

Unfortunately, this is *not* true for general linear transformations between normed vector spaces! The class of linear transformations for which this *is* true is called the class of *bounded linear operators*.

**Definition 1.27 (Bounded Linear Operator/Transformation)** Consider two normed vector spaces,  $(V, \|\cdot\|_V)$  and  $(W, \|\cdot\|_W)$ , and a linear transformation  $A : V \rightarrow W$ .  $A$  is said to be a bounded linear operator/transformation if there exists a  $K \geq 0$  for which

$$\|Ax\|_W \leq K \|x\|_V, \quad \forall x \in V. \quad (1.36)$$

*Remark 1.15* Here, we've undergone a little bit of a terminology change from “transformation” to “operator.” The “operator” terminology is more commonly used in functional analysis, while “transformation” is more often seen in linear algebra. These two terms are entirely interchangeable - for instance one may say “bounded linear operator” or “bounded linear transformation” in reference to the definition above.

Thus, a linear transformation is bounded if it doesn't “scale” any vector to be arbitrarily large! Taking a closer look at the definition of a bounded linear operator, one immediately notices a similarity to the definition of Lipschitz continuity. This observation leads to the following result.

**Proposition 1.17** *Bounded linear operators are Lipschitz continuous.*

**Proof** Consider a bounded linear operator  $A : V \rightarrow W$ . Such a map satisfies  $\|Ax\|_W \leq K \|x\|_V$ ,  $\forall x \in V$ , where  $K \geq 0$  is some fixed constant. By the axioms of a vector space, for all  $x, y \in V$ ,  $x - y \in V$ . Therefore, one has that for all  $x, y \in V$ ,

$$\|Ax - Ay\|_W = \|A(x - y)\|_W \leq K \|x - y\|_V. \quad (1.37)$$

We conclude that  $A$  is Lipschitz continuous with Lipschitz constant  $K$ .  $\square$

One of the most interesting and useful facts about the set of bounded linear operators between two vector spaces is that they themselves form a vector space! In the following theorem, we define the vector space of bounded linear operators.

**Theorem 1.2 (Vector Space of Bounded Linear Operators)** *Consider two normed vector spaces  $(V, \|\cdot\|_V)$  and  $(W, \|\cdot\|_W)$  over  $\mathbb{K}$ . The set of all bounded linear operators between  $V$  and  $W$ , denoted  $\mathcal{L}(V, W)$ , is itself a vector space under the operations  $+$  and  $(\cdot)$ , defined*

$$(A + B)(v) = A(v) + B(v) \quad (1.38)$$

$$(c \cdot A)(v) = c(Av), \quad (1.39)$$

for all  $v \in V$  and  $c \in \mathbb{K}$ .

**Exercise 1.25** Prove Theorem 1.2.

One may also verify that the *composition* of any two bounded linear operators is a bounded linear operator.

**Proposition 1.18 (Composition of Bounded Linear Operators)** *Consider two bounded linear operators,  $A \in \mathcal{L}(V, W)$  and  $B \in \mathcal{L}(U, V)$ . The composition of  $A$  and  $B$ , denoted  $AB$ , is defined  $ABv = A(Bv)$ .  $AB$  is a bounded linear operator from  $U$  to  $W$ .*

*Remark 1.16* The convention of writing the composition  $A \circ B$  as  $AB$  derives from matrix multiplication, where one writes the product of two matrices  $A$  and  $B$  as  $AB$ .

Now that we've defined a vector space of bounded linear operators, it's natural to seek out a norm that turns this vector space into a normed vector space. A natural choice would be to define a norm on  $\mathcal{L}(V, W)$  in terms of the norms on  $V$  and  $W$ .

**Definition 1.28 (Induced Operator Norm)** Let  $(V, \|\cdot\|_V)$  and  $(W, \|\cdot\|_W)$  be normed vector spaces and  $A \in \mathcal{L}(V, W)$  a bounded linear operator from  $V$  to  $W$ . The induced operator norm of  $A$  is defined,

$$\|A\|_{V,W} = \sup_{x \in V \setminus \{0\}} \frac{\|Ax\|_W}{\|x\|_V}. \quad (1.40)$$

*Remark 1.17* The “induced” in the name “induced operator norm” refers to the fact that the norm  $\|\cdot\|_{V,W}$  is induced by the norms on  $V$  and  $W$ . For convenience, one often refers to an induced operator norm simply as an “operator norm.” If no choice of norm on an operator is specified, it is typically assumed to be the operator norm induced by the normed vector spaces it maps between.

*Remark 1.18* The  $\sup_{x \in V \setminus \{0\}}$  simply ensures that we don't divide by zero in the quotient defining the operator norm. Shortly, we'll state a few equivalent ways of calculating the operator norm that avoid the use of a quotient.

An important special case of the operator norm is the following.

**Definition 1.29 (Left Multiplication Operator/Induced Matrix Norm)** Consider a matrix  $A \in \mathbb{K}^{m \times n}$ , and two normed vector spaces  $(\mathbb{K}^m, \|\cdot\|_b)$  and  $(\mathbb{K}^n, \|\cdot\|_a)$ .

1. **Left Multiplication Operator:** the linear operator of left multiplication with  $A$ , is the operator  $L_A : \mathbb{K}^n \rightarrow \mathbb{K}^m$ , defined  $L_A v := Av$  for all  $v \in \mathbb{K}^n$ .
2. **Induced Matrix Norm:** the induced matrix norm of  $A$  is defined  $\|A\|_{a,b} = \|L_A\|_{V,W}$ .

*Remark 1.19* If the spaces  $\mathbb{K}^m$  and  $\mathbb{K}^n$  have the same type of norm, for instance the 2-norm, one will simply denote the induced matrix norm as  $\|A\|_2$ , rather than  $\|A\|_{2,2}$ . In this case, one refers to the induced 2-norm of  $A$  simply as the 2-norm of  $A$ .

*Example 1.18 (Matrix 2-Norm)* The induced 2-norm of  $A \in \mathbb{K}^{m \times n}$  (where  $\mathbb{K} = \mathbb{R}$  or  $\mathbb{K} = \mathbb{C}$ ), is calculated  $\|A\|_2 = \sigma_{\max}(A)$ .

**Proposition 1.19 (Operator Norms Define a Normed Vector Space)** Let  $(V, \|\cdot\|_V)$  and  $(W, \|\cdot\|_W)$  be normed vector spaces. The space  $\mathcal{L}(V, W)$ , equipped with the operator norm  $\|\cdot\|_{V,W}$ , is a normed vector space.

**Exercise 1.26** Prove Proposition 1.19 by showing  $\|\cdot\|_{V,W}$  is a norm on  $\mathcal{L}(V, W)$ .

It's important to note that the operator norm  $\|\cdot\|_{V,W}$  is generally not the only norm giving  $\mathcal{L}(V, W)$  the structure of a normed vector space. Rather, operator norms are a *natural* choice of norm that derive from the normed vector spaces they map between.

*Example 1.19* The Frobenius norm of a matrix  $A \in \mathbb{K}^{m \times n}$ ,  $\|A\|_F = \sqrt{\text{tr}(A^*A)}$ , is not induced by any  $\ell^p$ -norms on  $\mathbb{K}^n$  and  $\mathbb{K}^m$ .

In addition to endowing the vector space  $\mathcal{L}(V, W)$  the structure of a normed vector space, the induced operator norm  $\|\cdot\|_{V,W}$  enjoys a number of other useful properties.

**Proposition 1.20 (Properties of the Operator Norm)** *Consider three normed vector spaces,  $(U, \|\cdot\|_U)$ ,  $(V, \|\cdot\|_V)$ , and  $(W, \|\cdot\|_W)$ . The following properties are satisfied:*

1. *Submultiplicative:* For all  $A \in \mathcal{L}(V, W)$  and  $B \in \mathcal{L}(U, V)$ ,  $\|AB\|_{U,W} = \|A\|_{V,W} \|B\|_{U,V}$ .
2. *Vector Inequality:* For all  $A \in \mathcal{L}(V, W)$  and  $x \in V$ ,  $\|Ax\|_W \leq \|A\|_{V,W} \|x\|_V$ .
3. *Equivalent Definitions:* The operator norm is equivalently computed by the formulas,

$$\|A\|_{V,W} = \sup_{\|x\|_V=1} \|Ax\|_W \text{ and } \|A\|_{V,W} = \inf\{K \geq 0 : \|Ax\| \leq K \|x\| \ \forall x \in V\}. \quad (1.41)$$

Let's run through the different components of this proposition. The first item of the proposition tells us that operator norms are submultiplicative - if we take two bounded linear operators and *compose* them, the operator norm of their composition  $AB : U \rightarrow W$  is bounded above by the product of their operator norms. The second item tells us that the operator norm provides us with an upper bound on how much a linear operator scales the norm of a vector. The final item tells us two equivalent ways of calculating the operator norm. Due to the scaling property of norms, one has

$$\sup_{\|x\|_V=1} \|Ax\|_W = \sup_{x \in V \setminus \{0\}} \frac{\|Ax\|_W}{\|x\|_V}. \quad (1.42)$$

This enables us to calculate the operator norm without the use of a quotient. The second formula for the operator norm - which is not as practical for computation - yields an interpretation of the operator norm as the *tightest possible* Lipschitz constant of the operator  $A$ . To complete our study of the space of linear operators, we define some special bounded linear operators.

**Definition 1.30 (Identity Operator)** Let  $V$  be a normed vector space. The identity operator on  $V$ ,  $\text{Id}_V : V \rightarrow V$  is defined  $\text{Id}_V v = v$ , for all  $v \in V$ .

**Exercise 1.27** Assuming both the domain and codomain instances of  $V$  are equipped with the same norm, confirm that the identity operator is a bounded linear operator.

Note that if  $V = \mathbb{R}^n$ , the identity linear operator  $\text{Id}_{\mathbb{R}^n}$  coincides with the linear operator  $L_{I_n}$  of left multiplication by the  $n \times n$  identity matrix,  $I_n$ .

**Definition 1.31 (Invertible Linear Operator)** Let  $V$  and  $W$  be normed vector spaces. A linear operator  $A \in \mathcal{L}(V, W)$  is said to be invertible if there exists a  $B \in \mathcal{L}(W, V)$  for which  $AB = \text{Id}_W$  and  $BA = \text{Id}_V$ . In this case,  $B$  is said to be the *inverse* of  $A$ , denoted  $A^{-1}$ .

*Remark 1.20* Note that we have used the language *the* inverse. One may show that the inverse of a linear operator is always unique - thus, it makes sense to talk about *the* inverse rather than *an* inverse.

Using invertible linear operators, one defines what it means for two vector spaces to be *isomorphic*.

**Definition 1.32 (Isomorphism)** Two normed vector spaces  $(V, \|\cdot\|_V)$  and  $(W, \|\cdot\|_W)$  are said to be isomorphic if there exists an invertible linear mapping  $A \in \mathcal{L}(V, W)$ . Such an  $A$  is said to be an isomorphism of the vector spaces  $V$  and  $W$ .

Roughly speaking, if two vector spaces are isomorphic, they share the same underlying “algebraic structure.” It is a fundamental theorem of linear algebra that every finite-dimensional vector space over  $\mathbb{R}$  is isomorphic to  $\mathbb{R}^n$  for some  $n$ . We conclude that any finite-dimensional vector space over  $\mathbb{R}$  has the same *algebraic structure* as  $\mathbb{R}^n$ , for some  $n$ . Consequently, it’s often sufficient to prove results for general finite-dimensional vector spaces in  $\mathbb{R}^n$ .

## 1.4 Banach Spaces

Thus far in our study of normed vector spaces, we’ve focused primarily on questions regarding the structure of sets, norms, and mappings of vector spaces. In this section, we study *convergence*, another fundamental subject of basic analysis. In order to fully understand convergence, we’ll need a little bit more structure than the basic normed vector space setting. In this section, we develop the basic theory of *Banach spaces*, a special class of normed vector space which enjoys additional convergence properties.

In order to define Banach spaces, we first need to understand sequences in normed vector spaces. Let’s start by discussing sequences in  $\mathbb{R}$ . In  $\mathbb{R}$ , we think of sequences as ordered lists of real numbers, for example,

$$(a_1, a_2, a_3, \dots), \text{ where each } a_i \in \mathbb{R}. \quad (1.43)$$

What such a sequence *really* is is a mapping from  $\mathbb{N} \rightarrow \mathbb{R}$ , taking an index of the sequence in  $\mathbb{N}$  and mapping it to a value in  $\mathbb{R}$ . Above, we map  $1 \mapsto a_1$ ,  $2 \mapsto a_2$ , and so on. This definition is generalized to the setting of vector spaces.

**Definition 1.33 (Sequence)** A sequence in a vector space  $V$  is a mapping  $a : \mathbb{N} \rightarrow V$ . Individual elements of the sequence are denoted  $a_n$ , while the entire sequence object is denoted  $\{a_n\} \subseteq V$ .

*Remark 1.21* The definition of a sequence of a mapping  $a : \mathbb{N} \rightarrow V$  suggests that we might write elements of a sequence as  $a(n)$  rather than  $a_n$ . However, as we often like to distinguish sequences from “regular” mappings, we favor the notation  $a_n$  over  $a(n)$ . Likewise, instead of referring to a sequence via the actual mapping  $a : \mathbb{N} \rightarrow V$ , one writes  $\{a_n\} \subseteq V$  to define a sequence  $a : \mathbb{N} \rightarrow V$ .

Using the language of normed vector spaces, we may formulate a precise definition for the *convergence* of a sequence.

**Definition 1.34 (Sequential Convergence)** Consider a normed vector space  $(V, \|\cdot\|)$ . A sequence  $\{a_n\} \subseteq V$  is said to converge if there exists a vector  $a \in V$  such that, for all  $\epsilon > 0$ , there exists an  $N \in \mathbb{N}$  (possibly dependent on  $\epsilon$ , for which  $n \geq N \implies \|a_n - a\| < \epsilon$ ). In this case, one says that  $a$  is the limit of  $\{a_n\}$ , and writes  $\lim_{n \rightarrow \infty} a_n = a$ .

What is this definition saying? Essentially, a sequence  $\{a_n\}$  converges to a limit  $a$  if  $a_n$  eventually comes and remains arbitrarily close to  $a$ . Here,  $\epsilon$  encodes a specification of how close we want the sequence to come to  $a$ , and  $N$  tells us how far into the sequence we need to look for  $\{a_n\}$  to come and remain within a distance  $\epsilon$  of  $a$ .

Since we’re working in normed vector spaces, sequences can take on far more interesting forms than simple sequences of real numbers. Sequences of *functions* will be of particular

interest to us. In the following example, we focus on sequences in a particularly important function space.

*Example 1.20 (Uniform Convergence)* Consider the normed vector space  $(V, \|\cdot\|_\infty)$ , where  $V$  is the set of functions  $f : \mathbb{R} \rightarrow \mathbb{R}$  with finite supremum norm,

$$\|f\|_\infty = \sup_{t \in \mathbb{R}} |f(t)|. \quad (1.44)$$

Sequences in this normed vector space are sequences of *functions*,  $\{f_n\} \subseteq V$ . If a sequence of functions  $\{f_n\}$  converges to a function  $f$  with respect to the supremum norm  $\|\cdot\|_\infty$ , one says that  $f_n$  *converges uniformly* to  $f$ .

Uniform convergence has a particularly useful relationship with differentiation.

**Theorem 1.3 (Uniform Convergence & Differentiation)** *Let  $I \subseteq \mathbb{R}$  be a compact interval. Suppose  $\{f_n\}$ ,  $f_n : I \rightarrow \mathbb{R}$ , is a sequence of functions for which  $\{f_n(t_0)\}$  converges for some  $t_0 \in I$ . If  $\{f'_n\}$  converges uniformly on  $I$ , then  $\{f_n\}$  converges uniformly on  $I$  to a differentiable function  $f$ , with  $f'(t) = \lim_{n \rightarrow \infty} f'_n(t)$  for all  $t \in I$ .*

**Proof** See [33] for the details. □

Fortunately, sequences interact well with the algebraic operations of a vector space, as well as with continuous functions between normed vector spaces.

**Proposition 1.21 (Sequential Limit Properties)** *Let  $(V, \|\cdot\|_V)$  and  $(W, \|\cdot\|_W)$  be normed vector spaces over  $\mathbb{K}$  and  $\{a_n\}, \{b_n\} \subseteq V$  convergent sequences with limits  $a$  and  $b \in V$ .*

1. Algebraic Combinations: for  $\alpha, \beta \in \mathbb{K}$ ,  $\lim_{n \rightarrow \infty} (\alpha a_n + \beta b_n) = \alpha a + \beta b$ .
2. Function Composition: for any continuous mapping  $f : V \rightarrow W$ ,  $\lim_{n \rightarrow \infty} f(a_n) = f(a)$ .

So far, we've discussed a definition of convergence which requires a candidate limit in order to certify convergence. Although in simple cases, it's not too hard to come up with a candidate for a limit (e.g. we can guess  $1/n \rightarrow 0$  without too much trouble), for more complex sequences this becomes challenging.

What we'd like is a way to certify a sequence converges *without* knowing ahead of time what the sequence converges to. Let's reason about how we might do this in  $\mathbb{R}$ . In  $\mathbb{R}$ , our intuition tells us that a sequence  $\{a_n\}$  will converge if its terms get closer and closer together as  $n$  grows large. We generalize this idea to normed vector spaces with the following definition.

**Definition 1.35 (Cauchy Sequence)** Let  $(V, \|\cdot\|)$  be a normed vector space. A sequence  $\{a_n\} \subseteq V$  is said to be a Cauchy sequence if, for all  $\epsilon > 0$ , there exists an  $N > 0$  (possibly dependent on  $\epsilon$ ), such that  $n, m \geq N \implies \|a_n - a_m\| < \epsilon$ .

*Remark 1.22* Frequently, instead of saying that a sequence  $\{a_n\} \subseteq V$  is a Cauchy sequence, we will simply say “ $\{a_n\}$  is Cauchy,” and drop the extra label of “sequence.”

Based on the definition, we see that a sequence is Cauchy if, given any  $\epsilon > 0$ , the terms of the sequence eventually come and remain within a distance  $\epsilon$  of each other. In other words, for large  $n$ , the terms of the sequence begin to “cluster” together. Notably, since this definition only relies on the terms of the sequence, one does not require a candidate limit to prove that a given sequence is Cauchy.

Let's determine if Cauchy sequences meet our requirement. Is it true that a Cauchy sequence in any normed vector space converges? In finite dimensions, the answer is *yes*, but in infinite dimensions, the answer is (of course) much more subtle. Since it is not a given that a Cauchy sequence converges in any given normed vector space, we define a special class of normed vector spaces in which Cauchy sequences *always* converge. Although this might initially seem like a restrictive condition, this class of normed vector spaces turns out to be quite vast, containing a wide variety of interesting spaces.

**Definition 1.36 (Banach Space)** A Banach space is a normed vector space  $(V, \|\cdot\|)$  in which every Cauchy sequence converges to a limit in  $V$ .

*Remark 1.23* Spaces in which Cauchy sequences always converge to a limit in the space are also referred to as *complete spaces* - this terminology extends to more general spaces beyond normed vector spaces. Using this language, one refers to a Banach space as a *complete* normed vector space.

*Remark 1.24* It's critical to note that for a space  $V$  to be a Banach space, the limits of all Cauchy sequences must belong to  $V$ . It's not enough to have a Cauchy sequence converge to a limit outside of the space  $V$  - the limit must be contained in  $V$  itself.

Since every Cauchy sequence in a Banach space converges, Banach spaces afford us the ability to study the convergence of sequences without actually knowing what the limits of the sequences might be. This is an incredibly powerful ability that has far reaching implications. In fact, one can use the consequences of completeness to prove the existence and uniqueness of solutions to certain differential equations.

Now, we consider some important examples of Banach spaces. As we alluded to above, it is true that *every* finite-dimensional normed vector space is a Banach space.

**Theorem 1.4** Any finite-dimensional normed vector space is a Banach space.

*Proof* See Problem 1.3. □

This result alone encompasses an enormous class of interesting spaces. Because of Theorem 1.4, we must turn to infinite-dimensional spaces to find other examples of Banach spaces. A particularly rich class of Banach spaces is supplied by the  $\ell^p$  and  $L^p$  spaces, which we now define.

**Definition 1.37 ( $\ell^p$  Space)** Fix a number  $p \in [1, \infty)$ . The normed vector space  $(\ell^p, \|\cdot\|_{\ell^p})$  has as vectors the set of all sequences  $u : \mathbb{N} \rightarrow \mathbb{R}$  with finite  $\ell^p$  norm,

$$\|u\|_{\ell^p} = \left( \sum_{n=1}^{\infty} |u_n|^p \right)^{\frac{1}{p}} < \infty. \quad (1.45)$$

This space is equipped with the operations of addition and scalar multiplication of sequences. For  $p = \infty$ , the normed vector space  $(\ell^\infty, \|\cdot\|_{\ell^\infty})$  consists of all sequences  $u : \mathbb{N} \rightarrow \mathbb{R}$  with finite  $\ell^\infty$  norm,

$$\|u\|_{\ell^\infty} = \sup_{n \in \mathbb{N}} |u_n| < \infty. \quad (1.46)$$

*Remark 1.25* The definition of an  $\ell^p$  space can be easily extended from sequences  $u : \mathbb{N} \rightarrow \mathbb{R}$  to sequences with other domains and codomains, for instance  $u : \mathbb{Z}_{\geq 0} \rightarrow \mathbb{R}$  or  $u : \mathbb{N} \rightarrow \mathbb{R}^m$ .

In these two examples, one would adjust the indices of the sum or trade the absolute value for a norm on  $\mathbb{R}^m$  to accommodate the different domain or codomain of  $u$ . We'll return to the general case in Chapter 3.

As one can observe from the definition, the vectors of an  $\ell^p$  space are *infinite sequences* of real numbers with bounded  $\ell^p$  norm.

**Definition 1.38 ( $L^p$  Space)** Fix a number  $p \in [1, \infty)$ . The normed vector space  $(L^p, \|\cdot\|_{L^p})$  consists of all functions  $f : \mathbb{R} \rightarrow \mathbb{R}$  with finite  $L^p$  norm,

$$\|f\|_{L^p} = \left( \int_{\mathbb{R}} |f(x)|^p dx \right)^{\frac{1}{p}} < \infty \quad (1.47)$$

This space is equipped with the operations of function addition and scalar multiplication. For  $p = \infty$ , the normed vector space  $(L^\infty, \|\cdot\|_{L^\infty})$  consists of all functions  $f : \mathbb{R} \rightarrow \mathbb{R}$  with finite  $L^\infty$  norm,

$$\|f\|_{L^\infty} = \sup_{x \in \mathbb{R}} |f(x)| < \infty. \quad (1.48)$$

As opposed to  $\ell^p$  spaces, where elements are real-valued sequences, in an  $L^p$  space, the elements are real-valued *functions* with bounded  $L^p$  norm. Despite the change from sequence to function, the  $\ell^p$  and  $L^p$  norms have clear parallels in their definitions.

*Remark 1.26* As with  $\ell^p$  spaces, the definition of  $L^p$  spaces provided above is easily extended to more general domains and codomains (for example to functions taking values in  $\mathbb{R}^n$ ). As with  $\ell^p$ , we'll return to the general case in Chapter 3.

*Remark 1.27* In defining  $L^p$  spaces as spaces of functions with finite integrals for  $p \in [1, \infty)$ , we've glossed over a number of concerns regarding Riemann/Lebesgue integrability of the functions  $f$ . Since properly treating these concerns requires a (significant & not immediately revealing) detour into a measure theory, we simply assume that elements of  $L^p$  spaces are “well-behaved” enough to have well-defined integrals. We also ignore the concerns that follow from functions differing on sets of measure zero having the same norm. We direct the interested reader to the additional reading specified at the end of the chapter for a rigorous treatment of these spaces.

In the previous section, we promised that we would return to finish the story of compactness once we defined Banach spaces. Now, we make good on this promise and provide a characterization of compactness in Banach space through convergent sequences. In order to state this characterization, we first require the concept of a *subsequence*. Given a sequence,

$$a_1, a_2, \dots, a_n, \dots \quad (1.49)$$

in a vector space  $V$ , a subsequence is a subset of the sequence in which elements are picked out in an order that respects that of the original sequence. For instance, a valid subsequence of the sequence above would be,

$$a_1, a_3, a_5, \dots \quad (1.50)$$

since the terms of the subsequence appear in the same order as the original sequence. An *invalid* subsequence would then be,



$$a_2, a_1, a_3, a_2, \dots, \quad (1.51)$$

which does *not* form a valid subsequence since the terms appear out of order compared to the original sequence. We abstractly define this “preservation of order” requirement in the following definition of a subsequence.

**Definition 1.39 (Subsequence)** Let  $V$  be a vector space and  $\{a_n\} \subseteq V$  a sequence. A subsequence of  $\{a_n\}$  is a subset  $\{a_{n_k}\} \subseteq \{a_n\}$ , where  $n_k : \mathbb{N} \rightarrow \mathbb{N}$  is a strictly increasing sequence of indices,

$$n_1 < n_2 < n_3 < \dots, \quad (1.52)$$

which specify the indices of the terms drawn from  $\{a_n\}$ .

For instance, for the subsequence  $a_1, a_3, a_5, \dots$ , one would define  $n_k$  by  $n_k = 2k - 1$ . This produces the subsequence,

$$a_{n_1} = a_1, a_{n_2} = a_3, a_{n_3} = a_5, \dots, \quad (1.53)$$

which is exactly the desired subsequence. Clearly, the indices of the subsequence are strictly increasing. Using the language of subsequences, one may formulate an equivalent definition of compactness in Banach spaces. Consider the following theorem, the proof of which is beyond our scope.

**Theorem 1.5 (Compactness in Banach Space)** Let  $(V, \|\cdot\|)$  be a Banach space and  $K \subseteq V$ . The following provide two equivalent characterizations of the compactness of  $K$ :

1.  $K$  is compact iff every sequence  $\{a_n\} \subseteq K$  has a subsequence with a limit in  $K$ .
2.  $K$  is compact iff every sequence  $\{a_n\} \subseteq K$  has a subsequence that is Cauchy.

*Remark 1.28* The abbreviation “*iff*” is commonly used as shorthand for “if and only if.”

Thus, in Banach spaces, one can detect the compactness of a set  $K$  knowing only information about the sequences contained in  $K$ .

**Exercise 1.28** Using Theorem 1.5, prove that a single-element subset  $\{v\}$  of a Banach space (called a *singleton set*) is compact.

## 1.5 A Refresher on ODEs

Ordinary differential equations (ODEs) are the lingua franca of control in continuous time. In this section, we review some basic properties of scalar, linear differential equations. We postpone a formal discussion of the existence and uniqueness of solutions to such equations until Chapter 2, and only touch upon the most essential conceptual aspects here. As such, our treatment is significantly more informal than the previous sections of this chapter. For the reader concerned by the appalling lack of theorems - don’t worry, plenty are coming down the pipeline - sit tight for a few more pages!

We begin by discussing initial value problems. Here, we’ll keep our discussion to the case of ordinary differential equations in  $\mathbb{R}$ . Let  $f : \mathbb{R} \rightarrow \mathbb{R}$  be a scalar function. The ordinary differential equation (ODE) governed by  $f$  is the equation,

$$\frac{d}{dt}x(t) = f(x(t)). \quad (1.54)$$

To find the *general solution* to the ordinary differential equation, one must identify all functions of the form  $x : I \rightarrow \mathbb{R}$ , where  $I \subseteq \mathbb{R}$  is a nonempty open interval, satisfying  $\frac{d}{dt}x(t) = f(x(t))$ , for all  $t \in I$ .

Since it's a bit cumbersome to repeatedly write  $\frac{d}{dt}$  one writes  $\frac{d}{dt}x(t)$  in shorthand as  $\dot{x}(t)$ . Additionally, it's common to suppress the argument of  $x(t)$ . We therefore write the ordinary differential equation in shorthand as,

$$\dot{x} = f(x). \quad (1.55)$$

Now, we consider an *initial value problem* associated to an ordinary differential equation. Given a constant  $x_0 \in \mathbb{R}$ , the goal of the initial value problem is to solve the problem,

$$\frac{d}{dt}x(t) = f(x(t)), \quad x(0) = x_0 \in \mathbb{R}. \quad (1.56)$$

That is, one wishes to find a curve  $x : I \rightarrow \mathbb{R}$  (for  $I$  a nonempty, open interval), where  $0 \in I$ , satisfying  $\frac{d}{dt}x(t) = f(x(t))$ , for all  $t \in I$  and  $x(0) = x_0$ . In our shorthand introduced above, one would abbreviate this problem as

$$\dot{x} = f(x), \quad x(0) = x_0. \quad (1.57)$$

For a general nonlinear function  $f$ , initial value problems are challenging, and at worst impossible, to solve. A special case which has a well-defined solution is the scalar, linear initial value problem.

**Theorem 1.6 (Scalar, Linear, First Order IVP)** *Let  $a \in \mathbb{R}$  be a fixed scalar. Consider the initial value problem,*

$$\dot{x} = ax, \quad x(0) = x_0, \quad (1.58)$$

*where  $x_0 \in \mathbb{R}$  is a fixed initial condition. The unique solution  $x : \mathbb{R} \rightarrow \mathbb{R}$  to this initial value problem is  $x(t) = e^{at}x_0$ .*

**Exercise 1.29** Verify that the solution proposed in Theorem 1.6 solves the initial value problem. You do not need to prove uniqueness - we'll return to this in Chapter 2.

Interestingly, this result states that, not only is  $x(t) = e^{at}x_0$  a solution to the proposed initial value problem, but it is the *only* solution! That is, there is *no other function* satisfying the initial value problem. Further, the solution is defined on *all* of  $\mathbb{R}$ , rather than on some bounded, open interval. These properties are *not* shared by every initial value problem.

*Example 1.21* Consider the initial value problem,  $\dot{x} = 2\sqrt{|x|}$ ,  $x(0) = 0$ . For all  $a \geq 0$ , the function

$$x(t) = \begin{cases} (t-a)^2 & t \geq a \\ 0 & t < a, \end{cases} \quad (1.59)$$

is a solution of the initial value problem. Thus, the initial value problem has an *infinite* number of solutions.

Thus far, we've only considered *first order*, scalar ordinary differential equations. We can easily extend the definition of a differential equation to one that involves higher derivatives. In the following setup, we denote by  $x^{(n)}(t)$  the  $n$ 'th derivative  $\frac{d^n x(t)}{dt^n}$ . Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ , and consider the ordinary differential equation,

$$x^{(n)}(t) = f(x(t), x^{(1)}(t), \dots, x^{(n-1)}(t)). \quad (1.60)$$

Here, a solution of the differential equation is a function  $x : I \rightarrow \mathbb{R}$  whose  $n$ 'th derivative equals the function  $f$  of its first  $n - 1$  derivatives (by convention, the 0'th derivative of  $x(t)$  is taken to be  $x(t)$  itself).

Do we need to construct an entirely new theory for higher order differential equations? Fortunately, by lifting our problem from a *scalar* differential equation to a *vector* differential equation, we can transform any  $n$ 'th order scalar differential equation into a system of  $n$  first order differential equations. Consider the following change of variables for the  $n$ 'th order initial value problem specified above. Define,

$$q_0(t) = x(t) \quad (1.61)$$

$$q_1(t) = \dot{x}(t) \quad (1.62)$$

$$q_2(t) = \ddot{x}(t) \quad (1.63)$$

$$\vdots \quad (1.64)$$

$$q_{n-1}(t) = x^{(n-1)}(t). \quad (1.65)$$

Differentiating each of the  $q$  variables, one has

$$\dot{q}_0(t) = \dot{x}(t) = q_1(t) \quad (1.66)$$

$$\dot{q}_1(t) = \ddot{x}(t) = q_2(t) \quad (1.67)$$

$$\dot{q}_2(t) = x^{(3)}(t) = q_3(t) \quad (1.68)$$

$$\vdots \quad (1.69)$$

$$\dot{q}_{n-1}(t) = x^{(n)}(t) = f(x(t), \dots, x^{(n-1)}(t)). \quad (1.70)$$

We recognize that we can rewrite  $f(x(t), \dots, x^{(n-1)}(t))$  as  $f(q_0(t), \dots, q_{n-1}(t))$ . Thus, the differential equation  $x^{(n)}(t) = f(x(t), \dots, x^{(n-1)}(t))$  can be rewritten as a *vector* differential equation in  $\mathbb{R}^n$ ,

$$\frac{d}{dt} \begin{bmatrix} q_0(t) \\ \vdots \\ q_{n-1}(t) \end{bmatrix} = \begin{bmatrix} q_1(t) \\ \vdots \\ f(q_0(t), \dots, q_{n-1}(t)) \end{bmatrix}. \quad (1.71)$$

Defining a vector  $q = (q_0, \dots, q_{n-1}) \in \mathbb{R}^n$  and a function  $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$  as  $F(q) = (q_1, \dots, f(q_0, \dots, q_{n-1}))$ , we compactly rewrite this differential equation in vector form as,

$$\dot{q} = F(q). \quad (1.72)$$

To solve the differential equation, we must identify a *vector* function  $q : I \rightarrow \mathbb{R}^n$  satisfying  $\dot{q}(t) = F(q(t))$  for all  $t \in I$ . To recover the solution to our original, scalar ODE, we the

extract the component function  $q_0(t)$  of  $q(t)$ , which by definition equals the solution  $x(t)$  to the original ODE.

Initial value problems are also defined similarly to the scalar case. In order to define an initial value problem, one specifies a *vector* initial condition,  $q_0 \in \mathbb{R}^n$ , to get the problem,

$$\dot{q} = F(q), \quad q(0) = q_0. \quad (1.73)$$

Just as in the scalar case, the solution to the vector initial value problem is a function  $q : I \rightarrow \mathbb{R}^n$  satisfying  $\dot{q}(t) = F(q(t))$  for all  $t \in I$  and  $q(0) = q_0$ . The transformation from an  $n$ 'th order scalar ODE into a first order vector ODE tells us that it's sufficient just to develop a theory for first order, vector ODEs to study general ODEs. We resume this story in the next chapter!

## 1.6 Further Reading

For an abstract treatment of linear algebra, we refer the reader to the texts *Linear Algebra* by Friedberg, Insel, & Spence [15], and *Linear Algebra Done Right* by Sheldon Axler [5]. For a user-friendly introduction to real analysis in  $\mathbb{R}$ , we recommend *Understanding Analysis* by Stephen Abbott [1]. For a similarly user-friendly treatment of analysis in normed vector spaces & Banach spaces, we refer the reader to *Measure, Integration, & Real Analysis* by Sheldon Axler [4]. Here, the reader can find proofs of a number of the concepts treated in this section, as well as a rigorous treatment of  $L^p$  spaces, enabled by measure theory.

## 1.7 Problems

**Problem 1.1 (Consequences of Norm Equivalence)** In this problem, we'll consider some simple consequences of norm equivalence. Consider a vector space  $V$  with two equivalent norms,  $\|\cdot\|_a$  and  $\|\cdot\|_b$ .

1. Show that a sequence  $\{v_n\} \subseteq V$  converges with respect to  $\|\cdot\|_a$  if and only if it converges with respect to  $\|\cdot\|_b$ .
2. Show that a mapping  $f : V \rightarrow V$  is continuous with respect to norm  $\|\cdot\|_a$  if and only if it is continuous with respect to  $\|\cdot\|_b$ . Does the same property hold for Lipschitz continuity?

**Problem 1.2 (Unbounded Linear Operators)** We know that every linear transformation between finite-dimensional normed vector spaces is bounded. In infinite dimensions, we're not quite so lucky! Product an example of an unbounded linear transformation from  $\ell^2 \rightarrow \ell^2$ . *Hint: The  $\ell^2$  norm is defined as an infinite series - think about some series that converge, and how they can be linear modified to no longer converge. It helps to work with the square of the  $\ell^2$  norm here.*

**Problem 1.3 ( $(\mathbb{R}^n, \|\cdot\|)$  is a Banach Space)** One may show that in  $\mathbb{R}$ , a sequence converges (with respect to the absolute value norm) if and only if it is Cauchy. That is,  $(\mathbb{R}, |\cdot|)$  is a Banach space. In this problem, we'll show that for any norm  $\|\cdot\|$  on  $\mathbb{R}^n$ ,  $(\mathbb{R}^n, \|\cdot\|)$  is a Banach space - this is a special case of the result that *every* finite-dimensional normed vector space is Banach.

1. Consider a sequence  $\{v_k\} \subseteq \mathbb{R}^n$ . Let  $\{v_k^i\} \subseteq \mathbb{R}$  represent the sequence formed from the  $i$ 'th components of each  $v_k \in \mathbb{R}^n$  (i.e.  $v_k = (v_k^1, v_k^2, \dots, v_k^n)$ ). Show that the sequence  $\{v_k\}$  converges to a vector  $v \in \mathbb{R}^n$  with respect to the  $\ell^\infty$  norm on  $\mathbb{R}^n$  if and only if each component sequence  $\{v_k^i\}$  converges to  $v^i$  in  $\mathbb{R}$ .
2. Show that  $\{v_k\} \subseteq \mathbb{R}^n$  is a Cauchy sequence with respect to the  $\ell^\infty$  norm on  $\mathbb{R}^n$  if and only if each component sequence  $\{v_k^i\}$  is Cauchy in  $\mathbb{R}$ .
3. Using norm equivalence on  $\mathbb{R}^n$ , show that for any norm on  $\mathbb{R}^n$ , a sequence is Cauchy if and only if it is convergent.

**Problem 1.4 (The Space of Polynomials)** Consider the set  $\mathcal{P}$  consisting of all polynomials (of all finite degrees)  $p : I \rightarrow \mathbb{R}$  on a compact interval  $I \subseteq \mathbb{R}$ .

1. Show that  $(\mathcal{P}, \|\cdot\|_\infty)$ , where  $\|\cdot\|_\infty$  is the sup norm,  $\|p\|_\infty = \sup_{t \in I} |p(t)|$ , is a normed vector space.
2. Is  $(\mathcal{P}, \|\cdot\|_\infty)$  a Banach space? Explain why or why not. *Hint: think about Taylor series.*

**Problem 1.5 (Systems of First Order Equations)**

1. Show that an  $n$ 'th order linear ODE,

$$\frac{d^n x}{dt^n} + a_{n-1} \frac{d^{n-1} x}{dt^{n-1}} + \dots + a_1 \frac{dx}{dt} + a_0 x = 0, \quad a_i \in \mathbb{R}, \quad (1.74)$$

can be rewritten as a system of  $n$ , first order differential equations of the form,

$$\dot{z} = Az, \quad (1.75)$$

where  $z \in \mathbb{R}^n$  and  $A \in \mathbb{R}^{n \times n}$ . This tells us that it's sufficient to examine *linear systems of first order ODEs* in order to reach conclusions about linear  $n$ 'th order ODEs.

2. Show that an  $n$ 'th order recurrence,

$$x[k+n] + a_{n-1}x[k+n-1] + \dots + a_1x[k+1] + a_0x[k] = 0, \quad (1.76)$$

can be rewritten as a system of  $n$ , first order recurrences of the form,

$$z[k+1] = Az[k], \quad (1.77)$$

where  $z \in \mathbb{R}^n$  and  $A \in \mathbb{R}^{n \times n}$ . This tells us that it's sufficient to examine *linear systems of first order recurrences* in order to reach conclusions about linear  $n$ 'th order recurrences.

**Problem 1.6 (The Structured Singular Value ★★)** The (complex) structured singular value is a function from the set of  $n \times n$  complex matrices to the reals that helps us understand the “gain” of matrices with structured uncertainty. The first step towards defining the structured singular value is to define a set of matrices  $\underline{\Delta} \subseteq \mathbb{C}^{n \times n}$ . Let  $r_1, \dots, r_S$  and  $m_1, \dots, m_F$  be positive integers for which  $\sum_{i=1}^S r_i + \sum_{j=1}^F m_j = n$ . Then, define a set  $\underline{\Delta} \subseteq \mathbb{C}^{n \times n}$  as

$$\underline{\Delta} := \{\text{blkdiag}(\delta_1 I_{r_1}, \dots, \delta_S I_{r_S}, \Delta_{S+1}, \dots, \Delta_{S+F}) : \delta_i \in \mathbb{C}, \Delta_{s+j} \in \mathbb{C}^{m_j \times m_j}\}, \quad (1.78)$$

where  $I_k$  represents the  $k \times k$  identity matrix. In short,  $\underline{\Delta}$  is the set of block diagonal matrices with *repeated scalar blocks* of dimensions  $r_i \times r_i$  (these are the blocks  $\delta_i I_{r_i}$ ) and *full blocks* of dimensions  $m_j \times m_j$  (these are the blocks  $\Delta_{S+j}$ ). Given a matrix  $M \in \mathbb{C}^{n \times n}$  and

a set  $\underline{\Delta} \subseteq \mathbb{C}^{n \times n}$  of the form above, one defines the structured singular value of  $M$ ,  $\mu_{\underline{\Delta}}(M)$ , as follows.

**Definition 1.40 (Structured Singular Value)** For  $M \in \mathbb{C}^{n \times n}$ ,  $\mu_{\underline{\Delta}}(M)$  is defined,

$$\mu_{\underline{\Delta}}(M) := \frac{1}{\inf\{\bar{\sigma}(\Delta) : \Delta \in \underline{\Delta} \text{ and } \det(I - M\Delta) = 0\}}, \quad (1.79)$$

unless no  $\Delta \in \underline{\Delta}$  makes  $I - M\Delta$  singular, in which case  $\mu_{\underline{\Delta}}(M) := 0$ .

Note that here, we use  $\bar{\sigma}(M)$  to denote the maximum singular value of  $M$ . Based on this definition,  $\mu_{\underline{\Delta}}(M)$  depends both on  $M$  and on the set  $\underline{\Delta}$ . Now, let's get started on our analysis of  $\mu_{\underline{\Delta}}$ ! *Note: In the following problems, you can assume for simplicity that one does not encounter the case where no  $\Delta$  makes  $I - M\Delta$  singular.*

1. Compute  $\mu_{\underline{\Delta}}(M)$  in the case where  $\underline{\Delta}$  is *unstructured*, i.e.  $\underline{\Delta} = \mathbb{C}^{n \times n}$ .
2. Recall that the spectral radius of a matrix  $M \in \mathbb{C}^{n \times n}$  is defined,

$$\rho(M) := \max_i |\lambda_i(M)|. \quad (1.80)$$

Define the set  $B_{\underline{\Delta}} = \{\Delta \in \underline{\Delta} : \bar{\sigma}(\Delta) \leq 1\}$ . Prove that the structured singular value can be calculated as the following function of spectral radius:

$$\mu_{\underline{\Delta}}(M) = \sup_{\Delta \in B_{\underline{\Delta}}} \rho(\Delta M). \quad (1.81)$$

Now, consider the special case where  $\underline{\Delta} = \{\delta I_n : \delta \in \mathbb{C}\}$ . In this case, show that  $\mu_{\underline{\Delta}}(M) = \rho(M)$ .

3. Let's consider some additional methods of computing  $\mu$ . Define the following subset of  $\mathbb{C}^{n \times n}$ :

$$\underline{D} = \{\text{blkdiag}(D_1, \dots, D_S, d_{S+1}I_{m_1}, \dots, d_{S+F}I_{m_F} : D_i \in \mathbb{C}^{r_i \times r_i}, D_i \succ 0, d_{S+j} \in \mathbb{R}_{>0}\}. \quad (1.82)$$

Prove that, for all  $D \in \underline{D}$ ,

$$\mu_{\underline{\Delta}}(M) = \mu_{\underline{\Delta}}(D^{\frac{1}{2}}MD^{-\frac{1}{2}}). \quad (1.83)$$

Then, show that,

$$\mu_{\underline{\Delta}}(M) \leq \inf_{D \in \underline{D}} \bar{\sigma}(D^{\frac{1}{2}}MD^{-\frac{1}{2}}). \quad (1.84)$$

4. Fix a matrix  $M \in \mathbb{C}^{n \times n}$ . For the set  $\underline{D}$  introduced in part (3), show that the following set is convex for each fixed  $\beta \in \mathbb{R}$ :

$$\{D \in \underline{D} : \bar{\sigma}(D^{\frac{1}{2}}MD^{-\frac{1}{2}}) \leq \beta\}. \quad (1.85)$$

*Hint: Rewrite as a linear matrix inequality. Such inequalities are amenable to implementation in convex optimization solvers!*

## Chapter 2

# Linear Dynamical Systems

In this chapter, we begin in earnest our study of systems and control theory. First, we introduce the main players in linear systems theory: linear dynamical systems and their state space representations. Following this, we study solutions to state space representations of linear systems, and see how the state space representations generate formal linear dynamical systems. Then, we'll move on to study linear systems from the I/O perspective, developing the theory of Laplace and  $\mathcal{Z}$ -transforms along the way. Let's begin!

### 2.1 Dynamical Systems & State Space Models

In order to develop a precise, mathematical theory of systems and control, it's vital that we understand what control systems actually are. Let's begin with a simple, motivating example. Suppose I have a shower with a handle I can use to control the water temperature. If I stick my hand into the water, I can gain some information as to whether the temperature is too high or too low, and I can adjust the position of the handle accordingly. I can continue to repeat this process - move handle, touch water, correct handle position - until the shower is at a temperature I want. If I continually take measurements and adjust the handle position, I can also adapt to unanticipated changes in the environment, such as my roommates washing dishes and reducing the supply of hot water. In other words, by adapting my actions according to measurements, I can become *robust* to changes in the shower environment.

Fundamentally, this is an example of a *feedback control system*, a system with an input (the position of the handle), a measurement (the temperature of the water that I estimate with my hand), a state (the *true* temperature of the water), internal dynamics (how the handle position affects the water temperature), and a description of time (the number of seconds that have passed). By incorporating feedback from the environment into my actions, I can *control* the system to reach a desired temperature.

This is the fundamental idea of feedback control: by measuring the environment, we can make informed decisions that enable us to control the state of the environment. Additionally, by taking repeated measurements, we can make decisions that adapt to unexpected events that might occur. Thus, in addition to affording us the ability of control over our environment, feedback yields the potential to be *robust* to uncertainty and disturbances in the environment.

### 2.1.1 Causal Input/Output Dynamical Systems

In order to develop the mathematical foundations of systems and control theory, we first need a precise definition for a *system*. Let's distill the most essential components of the shower system to gain some insight into the problem of determining a precise definition for an abstract, mathematical system. First, let's focus on the different objects making up the shower system. For simplicity, we'll assume that we have an infinite supply of hot water, and that the temperature of the shower is entirely determined by its current temperature, the time that has passed, and the history of shower handle positions. The key objects of this simple shower system are the following.

1. Time: we know the time of day at which we entered the shower, and the current time of day. We can represent both the entry time and the current time with a real number,  $t \in \mathbb{R}$ , corresponding to the number of seconds (or any other appropriate unit of time) that have passed since the beginning of the day.
2. Inputs: the shower handle was an *input* to the shower system. Input signals to the system composed of different positions of the shower handle over time. We can measure the values of inputs by the angle,  $\theta(t) \in \mathbb{R}$ , of the shower handle at time  $t$ . An input signal would therefore be a function of time,  $\theta(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$ , assigning to each time a shower handle position.
3. Outputs: the outputs (measurements) we took of the shower system were tests of the water temperature with our hand. Since we only took measurements with our hand, and not a thermometer, the measurements of our system might take on values in a set,

$$\{\text{icy, cold, mild, hot, ouch!}\} \quad (2.1)$$

Measurement *signals* would then be mappings from time,  $t \in \mathbb{R}$ , to this set of measurement values.

4. State: we know that our set of measurements,  $\{\text{icy, cold, mild, hot, ouch!}\}$  doesn't quite cover the actual temperature of the system! We define the *state* of the shower to be the *actual* water temperature,  $T \in \mathbb{R}$ , measured in degrees Celsius (or any other appropriate unit of temperature). This describes the entire state of the shower at a given time  $t$ . Using knowledge of the state,  $T$ , time,  $t$ , and input, we should be able to *completely determine* what the shower will do next (within the scope of our very simple shower model).

How are all of these basic objects tied together? Underneath the shower system, we know there exist some *shower dynamics* that determine how the temperature of the shower changes according to the passage of time and the history of shower handle inputs. Additionally, we know there is some underlying map that determines which measurement value out of the set  $\{\text{icy, cold, mild, hot, ouch!}\}$  we will feel given any true temperature and time. These concepts are encoded in the following two maps.

1. State Transition Map: the state transition map of the shower determines how the state (true temperature) of the shower is influenced by the start time, current time, starting temperature of the water, and history of shower handle positions. This gives us a *complete description* of how the true temperature of the shower changes over time. Notably, the current state of the system only depends on the previous and current shower handle positions! The state does not depend on the future inputs to the system (our shower is unfortunately not fancy enough to predict the future).



2. Readout Map: given any time, temperature, and input value, we should know what measurement value our hand is feeling. The readout map maps from a pair of time, temperature, and shower handle position to this measurement value. Notice that the readout map is *memoryless* - it does not require a history of temperatures or a full input signal, only the current temperature and the current input value!

Finally, we know there are a couple of simple properties all of these objects should obey.

1. Time: we can add and subtract time in the shower system without any confusion.
2. Restriction: suppose we have two input signals to the system which match from times  $t_0$  to  $t_1$ . Over the time period  $t_0$  to  $t_1$ , these two input signals should produce the same behavior, regardless of if they differ after time  $t_1$ .
3. Composition: suppose over the time period  $t_0$  to  $t_1$ , an input signal takes us from temperature  $T_0$  to temperature  $T_1$ , and over the time period  $t_1$  to  $t_2$ , an input signal takes us from temperature  $T_1$  to temperature  $T_2$ . Then, applying the input signals from  $t_0$  to  $t_2$  should take us from  $T_0$  to  $T_2$ .
4. Identity: if no time passes, the temperature of the shower should stay the same.

When stated in context of the shower example, these three conditions are all fairly “obvious.” Although it may seem like these points are too trivial to mention, they’ll help us make a well-posed, abstract definition of a system that behaves in the way we expect.

This exploratory example provides us with a template definition for a formal *causal input/output dynamical system*. In order to state the formal definition, all we need to do is abstract away the details of the shower into the language of mathematics. As you’re reading the definition, relate the formal mathematical expressions to the analogous components of the shower system we outlined above.

Note that, in our formal definition, everything is named similarly to the shower example with the one exception: the “composition” property has been given the shiny new name of the “semigroup axiom” - we’ll discuss the rationale behind this after stating the definition.

**Definition 2.1 (Causal Input/Output Dynamical System)** Let  $\mathcal{T} \subseteq \mathbb{R}$  be a nonempty set. A causal, input/output dynamical system on  $\mathcal{T}$  is a tuple  $\mathcal{D} = (\mathcal{U}, \mathcal{Y}, \Sigma, \varphi, r)$ . Each term is defined as follows:

1. Time set:  $\mathcal{T}$  is the time set, a subset of  $\mathbb{R}$  describing the possible times in the system.
2. Input space:  $\mathcal{U}$  is the input space, a set of mappings from  $\mathcal{T}$  to a fixed set  $U$ ,

$$\mathcal{U} \subseteq \{u : \mathcal{T} \rightarrow U\}. \quad (2.2)$$

$U$  is referred to as the input value space. Elements of  $\mathcal{U}$  represent the possible input signals to the system.

3. Output space:  $\mathcal{Y}$  is the output space, a set of mappings from  $\mathcal{T}$  to a fixed set  $Y$ ,

$$\mathcal{Y} \subseteq \{y : \mathcal{T} \rightarrow Y\}. \quad (2.3)$$

$Y$  is referred to as the output value space. Elements of  $\mathcal{Y}$  represent the possible output (measurement) signals of the system.

4. State space:  $\Sigma$  is the state space, a set representing the possible states of the system.
5. State transition map:  $\varphi$  is the state transition map, a map,

$$\varphi : \mathbf{T} \times \Sigma \times \mathcal{U} \rightarrow \Sigma, \quad (2.4)$$

where  $\mathbf{T} = \{(t_1, t_0) \in \mathcal{T} \times \mathcal{T} : t_1 \geq t_0\}$ , which describes how the state of the system evolves. In particular, for  $t_0, t_1 \in \mathcal{T}$  with  $t_0 \leq t_1$ ,  $x_0 \in \Sigma$ , and  $u(\cdot) \in \mathcal{U}$ ,  $\varphi(t_1, t_0, x_0, u(\cdot))$  returns the state of the system at time  $t_1$  after starting from  $x_0$  at time  $t_0$  and applying input signal  $u(\cdot)$ .

6. Readout map:  $r$  is the readout map, a map,

$$r : \mathcal{T} \times \Sigma \times U \rightarrow Y, \quad (2.5)$$

which returns the measured output at time  $t$  given the system has a current state of  $x(t) \in \Sigma$  and a current input value of  $u(t) \in U$ .

A dynamical system  $\mathcal{D}$  must additionally satisfy the following four axioms:

1. Time axiom: for all  $t_1, t_2 \in \mathcal{T}$ ,  $t_1 + t_2 \in \mathcal{T}$  and  $t_1 - t_2 \in \mathcal{T}$ .
2. Restriction axiom: for all  $t_0, t_1 \in \mathcal{T}$  with  $t_0 \leq t_1$ ,  $x_0 \in \Sigma$ , and  $u_1(\cdot), u_2(\cdot) \in \mathcal{U}$ , one has that

$$u_1(t) = u_2(t) \quad \forall t \in [t_0, t_1] \cap \mathcal{T} \implies \varphi(t_1, t_0, x_0, u_1(\cdot)) = \varphi(t_1, t_0, x_0, u_2(\cdot)). \quad (2.6)$$

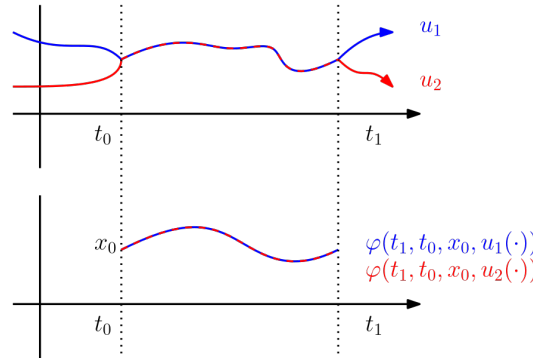
3. Semigroup axiom: for all  $t_0, t_1, t_2 \in \mathcal{T}$  with  $t_0 \leq t_1 \leq t_2$ ,  $x_0 \in \Sigma$ , and  $u(\cdot) \in \mathcal{U}$ ,

$$\varphi(t_2, t_1, \varphi(t_1, t_0, x_0, u(\cdot)), u(\cdot)) = \varphi(t_2, t_0, x_0, u(\cdot)). \quad (2.7)$$

4. Identity axiom: for all  $t \in \mathcal{T}$ ,  $x \in \Sigma$ , and  $u(\cdot) \in \mathcal{U}$ ,

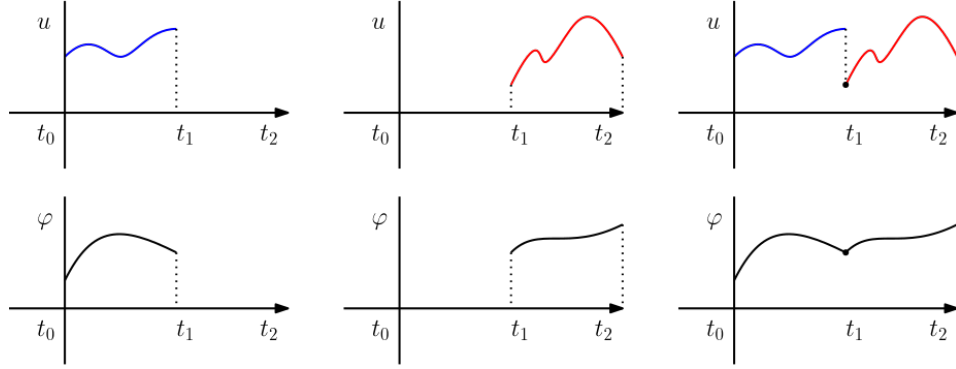
$$\varphi(t, t, x, u(\cdot)) = x. \quad (2.8)$$

Phew, what a mouthful! Let's highlight some subtle yet important consequences of the definition.



**Fig. 2.1** The restriction axiom states that if two input signals are equal on a time interval, they will produce the same state behavior on that time interval.

*Remark 2.1 (Signal versus Value)* In Definition 2.1, we define the input space  $\mathcal{U}$  and the output space  $\mathcal{Y}$  to be spaces of *signals*, not spaces of *values*! Associated to the space of input signals  $\mathcal{U}$ , we have the space of input values,  $U$ . Likewise, associated to the space of output signals  $\mathcal{Y}$ , we have the space of output values,  $Y$ . To make this concrete, the space of input



**Fig. 2.2** The semigroup axiom states that the state transition map is well-behaved under *composition*. The semigroup axiom implies that, if we stitch together two inputs, the resulting behavior is the same as that which results from applying the first input over its domain and the second input over its domain.

values might be  $U = \mathbb{R}^m$ , but the space of input signals might be the set of continuous maps from  $\mathbb{R} \rightarrow \mathbb{R}^m$ . To distinguish between the two, we will write  $u(\cdot)$  to represent a signal and  $u(t)$  to represent its value at time  $t$ .

*Remark 2.2 (Causality)* A system is said to be *causal* if its state and output behavior depends only on its previous and current inputs, *not* on its future inputs! As we can see from the definition of the state transition map, causality is *baked into* the definition of a causal dynamical system. The restriction axiom tells us that the state at time  $t_1$  *only* depends on the state  $x_0$  at time  $t_0$  and the input signal  $u(\cdot)$  applied from  $t_0$  to  $t_1$ . Thus, any values of the input signal after time  $t_1$  are entirely irrelevant to the behavior of the system. We conclude that any dynamical system satisfying Definition 2.1 must be causal.

*Remark 2.3 (Time Axiom)* The time axiom can also be stated in terms of *algebraic* language. One may equivalently require  $\mathcal{T}$  to be a *subgroup* of  $(\mathbb{R}, +)$ , the group of real numbers with the addition operation. If you're unfamiliar with algebraic language, this connection isn't something you need to worry about - there are no practical differences between this and what we stated in Definition 2.1.

*Remark 2.4 (Semigroup Axiom)* Why rename the “composition” axiom as the semigroup axiom? The name *semigroup* alludes to another connection between causal I/O dynamical systems and abstract algebra. If you're interested, read the definition of a semigroup and a semigroup action and see if you can draw a connection between a algebraic semigroups/semigroup actions and the semigroup axiom of Definition 2.1.

*Remark 2.5 (Generality)* This is but one of many different definitions of an I/O dynamical system, and is by no means the most general definition possible. For instance, the systems proposed above are causal, deterministic (not random), and have fixed input and readout spaces that do not change with state or time. Further, the state transition map is defined on the entire time set, rather than on subsets thereof. Although these assumptions are sufficiently general for our purposes in this course, one should keep these limitations in mind! We direct the reader to the references at the end of the chapter for the more general definitions.

To get some practice with identifying the components of a dynamical system, try the following three exercises. Make sure to state your assumptions where necessary; in each case, you'll need to lay out some simplifying assumptions in order to come up with a manageable dynamical system.

**Exercise 2.1** Come up with a dynamical system representing a falling rock. Specify each component of the tuple  $\mathcal{D} = (\mathcal{U}, \mathcal{Y}, \Sigma, \varphi, r)$ , as well as the time set  $\mathcal{T}$ . Explain why the time, restriction, semigroup, and identity axioms hold for this system.

**Exercise 2.2** Come up with a dynamical system representing an airplane. Specify each component of the tuple  $\mathcal{D} = (\mathcal{U}, \mathcal{Y}, \Sigma, \varphi, r)$ , as well as the time set  $\mathcal{T}$ . Explain why the time, restriction, semigroup, and identity axioms hold for this system.

**Exercise 2.3** Come up with a dynamical system representing a computer. Specify each component of the tuple  $\mathcal{D} = (\mathcal{U}, \mathcal{Y}, \Sigma, \varphi, r)$ , as well as the time set  $\mathcal{T}$ . Explain why the time, restriction, semigroup, and identity axioms hold for this system.

These three examples - of a falling rock, an airplane, and a computer - illustrate the flexibility of the dynamical system definition we posed above. Each system is vastly different, yet fits into the same framework under minimal assumptions.

However, this flexibility comes at a price. Whenever one makes a highly abstract, general definition such as Definition 2.1, there is a fundamental tradeoff that one makes. Almost always, generality comes at the expense of *practicality* - the more general the definition, the less practical it is, and the harder it is to come to interesting conclusions. In order to state more interesting results about the behavior of dynamical systems, we'll need to consider dynamical systems with more structure than the basic scaffolding offered by Definition 2.1.

First, we outline two simple classes of dynamical system, based on the examples of an airplane and a computer. In order to properly describe the behavior of an airplane, one needs to use a time set  $\mathcal{T} = \mathbb{R}$ . On the other hand, for a computer—which makes decisions at discrete instants—one might use the time set  $\mathcal{T} = \mathbb{Z}$ . We distinguish between systems on these two important time sets as follows.

**Definition 2.2 (Continuous-Time System)** A dynamical system  $\mathcal{D}$  is said to be a continuous-time system if its time set is  $\mathcal{T} = \mathbb{R}$ .

**Definition 2.3 (Discrete-Time System)** A dynamical system  $\mathcal{D}$  is said to be a discrete-time system if its time set is  $\mathcal{T} = \mathbb{Z}$ .

As a general rule of thumb, if the state of a system varies as time passes in seconds (with no fixed jumps or increments in time), the system will be continuous-time. On the other hand, if the state of a system jumps at fixed, discrete increments, the system will be discrete-time<sup>1</sup>.

These two classes of system provide some nice, additional structure to Definition 2.1. What other structure might be interesting to add? Let's think back to the example of a falling rock, and see if anything else jumps out at us that isn't explicitly covered by Definition 2.1.

Suppose the state of the rock is its position and velocity in space and that its output is its position. As an input, let's take a force acting on the rock. This enables us to treat the action of “dropping” the rock as an input signal. To track the trajectory of the rock after

<sup>1</sup> Note that the choice of  $\mathcal{T} = \mathbb{Z}$  for discrete-time is somewhat arbitrary - one could reasonably replace  $\mathbb{Z}$  with another countable subset of  $\mathbb{R}$  that satisfies the time axiom.

we drop it, we could check the readout map. This tells us that, if we drop the rock at time  $t_0$  from initial state  $x_0$ , the rock position at time  $t$  is,

$$\text{rock position at time } t = r(t, \varphi(t, t_0, x_0, u(\cdot)), u(t)). \quad (2.9)$$

What if, instead of dropping the rock at time  $t_0$ , we sat around for ten minutes and dropped the rock at time  $t'_0 = t_0 + 10$ ? Would we expect the rock to fall in the same way? If our rock is friends with Isaac Newton, the answer is—of course! To determine the trajectory of the rock, we shouldn't need to know the number of seconds since the beginning of time at which we drop it—what matters is how much time has *passed* since we dropped it. This behavior—of time *passed* being the relevant quantity of time—is prevalent in a number of systems. A dynamical system possessing this property is said to be *time-invariant*.

Before we can properly define time invariance, we must make a few auxiliary definitions. First, we make a definition that will reduce our notational overhead. We note that in the example of the falling rock, we composed the readout map with the state transition map in order to get the output at time  $t$ , given an initial time, initial state, and input signal. Since it's quite cumbersome to rewrite this composition every time we're interested in these objects, we define the following map.

**Definition 2.4 (Input/Output Map)** Given a dynamical system  $\mathcal{D}$ , the input/output map (I/O map) is the map  $\rho : \mathbf{T} \times \Sigma \times \mathcal{U} \rightarrow Y$  (for  $\mathbf{T} = \{(t_1, t_0) \in \mathcal{T} \times \mathcal{T} : t_1 \geq t_0\}$ ) defined as the composition,

$$\rho(t_1, t_0, x_0, u(\cdot)) = r(t_1, \varphi(t_1, t_0, x_0, u(\cdot)), u(t_1)), \quad (2.10)$$

of the readout and the state transition maps.

*Remark 2.6* An I/O map takes in a pair of times, an initial state, and an input signal and returns an output *value*, not an output signal! This is because we want the I/O map to convey information about the output at a particular time, rather than at all times.

With this definition made, we turn our attention back to the problem of defining time-invariance. First, we focus on how the components of a dynamical system change over time. We define a *delay-invariant* set of signals.

**Definition 2.5 (Delay-Invariant Set)** Consider a set of signals,  $\mathcal{U} \subseteq \{u : \mathcal{T} \rightarrow U\}$ , where  $\mathcal{T} \subseteq \mathbb{R}$  is a time set and  $U$  is an arbitrary set. If, for all  $\tau \in \mathcal{T}$  and all  $u(\cdot) \in \mathcal{U}$ , the signal

$$\hat{u} : \mathcal{T} \rightarrow U, \hat{u}(t) = u(t - \tau), \quad (2.11)$$

also belongs to  $\mathcal{U}$ , then  $\mathcal{U}$  is said to be a *delay-invariant set* with respect to  $\mathcal{T}$ .

*Remark 2.7* When the time set  $\mathcal{T}$  is clear from context, one refers to a delay-invariant set with respect to  $\mathcal{T}$  simply as a “delay-invariant set.” The “with respect to  $\mathcal{T}$ ” can be dropped.

*Remark 2.8* Delay-invariant sets are so-called since they are defined by delaying signals by a time  $\tau$ . It's important to note that if  $\tau < 0$ , a “delay” will actually shift a signal *forward* in time rather than shifting it backward. In this context, the name “delay” is therefore not entirely consistent with our intuitive understanding of the word.

*Remark 2.9* In this definition, we implicitly make use of the time axiom. Without the guarantee that  $t_1 - t_2 \in \mathcal{T} \forall t_1, t_2 \in \mathcal{T}$ , we would not know  $\hat{u}(t) = u(t - \tau)$  was a valid signal.

Thus, a set is delay-invariant if any signal in  $\mathcal{U}$  can be delayed by any time  $\tau$  and remain in  $\mathcal{U}$ . Equipped with this definition, we define a *delay map* on a delay-invariant set of signals.

**Definition 2.6 (Delay Map)** Consider a delay-invariant set of signals,  $\mathcal{U}$ . For  $\tau \in \mathcal{T}$ , the map  $T_\tau : \mathcal{U} \rightarrow \mathcal{U}$ , defined  $(T_\tau(u))(t) = u(t - \tau) \forall t \in \mathcal{T}$ , is called the delay map of time  $\tau$ .

Based on this definition, a shift map  $T_\tau$  simply *delays* any input signal by a fixed time  $\tau$ . Notice how the definition of a delay-invariant set ensures the delay map is well-defined—since we don't have to worry about a delayed signal leaving the set  $\mathcal{U}$ , we define the delay map to be a map from  $\mathcal{U} \rightarrow \mathcal{U}$ .

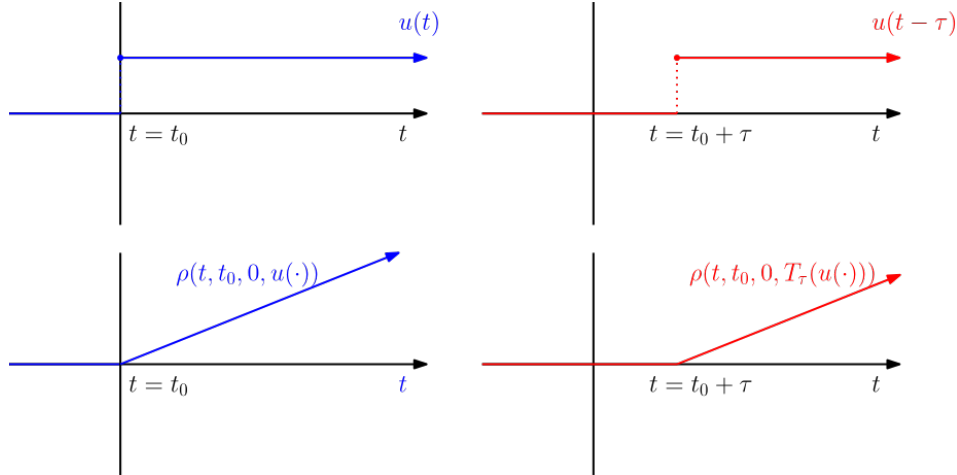
**Definition 2.7 (Time-Invariant System)** A causal I/O dynamical system  $\mathcal{D}$  is said to be time invariant if, for all  $\tau \in \mathcal{T}$ , the following conditions are satisfied:

1. Delay-invariant input space:  $\mathcal{U}$  is a delay-invariant set.
2. Delay-invariant output space:  $\mathcal{Y}$  is a delay-invariant set.
3. Delay-invariant transition map: For all  $t_0, t_1, \tau \in \mathcal{T}$  with  $t_0 \leq t_1$  and all  $x_0 \in \Sigma, u(\cdot) \in \mathcal{U}$ ,

$$\rho(t_1, t_0, x_0, u(\cdot)) = \rho(t_1 + \tau, t_0 + \tau, x_0, T_\tau(u(\cdot))), \quad (2.12)$$

where  $T_\tau : \mathcal{U} \rightarrow \mathcal{U}$  is the delay map of time  $\tau$  on  $\mathcal{U}$  and  $\rho$  is the I/O map of  $\mathcal{D}$ .

Item (3) of this definition—delay-invariant transition map—is by far the most important component of the definition. It states that the output of the system depends on the amount of time that has *passed*, rather than on the explicit start and end times. In particular, if we delay the inputs to the system by time  $\tau$ , we will get the same output at time  $t + \tau$  as the undelayed system at time  $t$ .



**Fig. 2.3** An example of a time-invariant response. On the left-hand side, we apply an input which jumps up at time  $t = t_0$ . The system responds by increasing along a ramp at time  $t = t_0$ . If we *delay* the input by  $\tau$ , the appearance of the ramp also delays by  $\tau$ . Thus, we observe that, for this input, the system respects the equality  $\rho(t + \tau, t_0 + \tau, T_\tau(u(\cdot))) = \rho(t, t_0, u(\cdot))$ .

What other structure can we add to a causal I/O dynamical system? Thus far, we've only placed structure on the *time* component of the I/O map, and haven't considered any algebraic or analytic conditions.

Although algebraically and analytically unstructured maps lend themselves well to generality, one cannot say the same for practicality. Without placing further algebraic or analytic constraints on the I/O map, we'll find it hard to perform any meaningful system analysis. For the class of *linear* I/O systems, however, a wide array of concepts become mathematically and computationally tractable. This is the class of systems we will focus on in this course.

**Definition 2.8 (Linear I/O System)** Consider an I/O system  $\mathcal{D} = (\mathcal{U}, \mathcal{Y}, \Sigma, \varphi, r)$  with I/O map  $\rho$ .  $\mathcal{D}$  is said to be a linear I/O system if the following conditions are satisfied:

1. Linear Spaces:  $\mathcal{U}, \mathcal{Y}$ , and  $\Sigma$  are vector spaces over a common field,  $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$ .
2. Linear I/O Map: For each fixed pair  $t_0, t_1 \in \mathcal{T}$  with  $t_0 \leq t_1$ , the I/O map is linear in  $\Sigma \times \mathcal{U}$ . That is, for all  $x_0, \hat{x}_0 \in \Sigma$ ,  $u(\cdot), \hat{u}(\cdot) \in \mathcal{U}$ , and  $\alpha, \beta \in \mathbb{K}$ ,

$$\rho(t_1, t_0, \alpha x_0 + \beta \hat{x}_0, \alpha u(\cdot) + \beta \hat{u}(\cdot)) = \alpha \rho(t_1, t_0, x_0, u(\cdot)) + \beta \rho(t_1, t_0, \hat{x}_0, \hat{u}(\cdot)). \quad (2.13)$$

*Remark 2.10* If the spaces  $\mathcal{U}, \mathcal{Y}, \Sigma$  of  $\mathcal{D}$  are all over a field  $\mathbb{K}$ , one says that  $\mathcal{D}$  itself is a system over a field  $\mathbb{K}$ .

*Remark 2.11* We'll refer to a linear I/O system simply as a "linear system" or a "linear dynamical system" where context allows.

Let's discuss the different components of the definition. The first condition, *linear spaces*, states that the input and output *signal* spaces are vector spaces, as is the state space  $\Sigma$ . This means that any linear combination of input signals,  $\alpha u(\cdot) + \beta \hat{u}(\cdot)$ , remains in the input space. Likewise, linear combinations of output signals and states remain in the output and state spaces, respectively.

The second condition, *linear I/O map*, states that the output of a linear I/O system must be linear in its initial condition and input. That is, if we scale the initial condition and output by the same value, the output should scale by that value as well. Additionally, if we add two sets of initial conditions and inputs, the corresponding output should be the sum of the individual outputs.

Let's get a basic feel for what this linear structure enables. In the following proposition, we state a few simple consequences of Definition 2.8.

**Proposition 2.1 (Output Response of Linear I/O Systems)** Any linear I/O system  $\mathcal{D}$  over a field  $\mathbb{K}$  satisfies the following:

1. Zero Input Response: For all  $x_0, \hat{x}_0 \in \Sigma$ ,  $t_0, t_1 \in \mathcal{T}$  with  $t_0 \leq t_1$ , and  $\alpha, \beta \in \mathbb{K}$ ,

$$\rho(t_1, t_0, \alpha x_0 + \beta \hat{x}_0, 0) = \alpha \rho(t_1, t_0, x_0, 0) + \beta \rho(t_1, t_0, \hat{x}_0, 0). \quad (2.14)$$

2. Zero State Response: For all  $u(\cdot), \hat{u}(\cdot) \in \mathcal{U}$ ,  $t_0, t_1 \in \mathcal{T}$  with  $t_0 \leq t_1$ , and  $\alpha, \beta \in \mathbb{K}$ ,

$$\rho(t_1, t_0, 0, \alpha u(\cdot) + \beta \hat{u}(\cdot)) = \alpha \rho(t_1, t_0, 0, u(\cdot)) + \beta \rho(t_1, t_0, 0, \hat{u}(\cdot)). \quad (2.15)$$

3. Zero Input/Zero State Decomposition: For all  $x_0 \in \Sigma$ ,  $u(\cdot) \in \mathcal{U}$ , and  $t_0, t_1 \in \mathcal{T}$  with  $t_0 \leq t_1$ ,

$$\rho(t_1, t_0, x_0, u(\cdot)) = \rho(t_1, t_0, x_0, 0) + \rho(t_1, t_0, 0, u(\cdot)). \quad (2.16)$$

Here,  $\rho(t_1, t_0, x_0, 0)$  is called the *zero-input response* and  $\rho(t_1, t_0, 0, u(\cdot))$  the *zero-state response*.

The last item of Proposition 2.1 tells us that, in order to understand the response of a linear I/O system to *any* input, all we need is the zero-input response and the zero-state response—there is an exact decomposition of the total response into the zero-input and zero-state components.

**Exercise 2.4** Prove Proposition 2.1.

Using our earlier definition of time-invariance, one may determine a further classification of linear dynamical systems.

**Definition 2.9 (Linear Time-Invariant/Varying System)** An I/O system  $\mathcal{D}$  is said to be linear, time-invariant (LTI) if it is a linear I/O system and it is time-invariant. If a linear I/O system is not necessarily linear, time-invariant, it is said to be linear, time-varying (LTV).

### 2.1.2 State Space Representations of Linear Systems

Thus far, we've only dealt with *abstract* dynamical systems, in which the evolution of the system is described by an arbitrary state transition map. Is this the most convenient way of describing a dynamical system? In practice, dynamical systems are typically not specified via their state transition map. Instead, one often specifies a set of equations (such as a differential equation or a recurrence relation) from which a state transition map can be determined. If we wish to describe a Newtonian physical system, for instance, we might start by writing down Newton's second law,  $F = m\ddot{x}$ , and deriving *differential equations* of motion. How do systems with dynamics of this form correspond to the abstract dynamical systems we discussed above?

In order to establish this connection, we make the important distinction between a *representation* of a dynamical system and the dynamical system itself. When we write down a differential equation such as  $F = m\ddot{x}$ , we define a differential equation *representation* of an abstract dynamical system. More generally, we say that a *system representation* is a description of a dynamical system using some mathematical framework (such as an ordinary differential equation, partial differential equation, recurrence relation, etc.) that fully determines the dynamical system.

**Definition 2.10 ((Informal) I/O System Representation)** An I/O system representation is a collection of mathematical data that uniquely determines a causal I/O dynamical system.

Let's return to the example of a physical system described by  $F = m\ddot{x}$  to illustrate what we mean by this. Let's take  $F$  to be the input to the dynamical system,  $(x, \dot{x})$  to be the state, and  $x$  to be the output. The pair,

$$\ddot{x} = \frac{1}{m}F, \quad r(x, \dot{x}, F) = x, \quad (2.17)$$

of an ordinary differential equation  $\ddot{x} = \frac{1}{m}F$  and an readout function  $r(x, \dot{x}, F) = x$ , together with sets of admissible inputs and outputs, constitute a representation of the abstract dynamical system. The solutions of the differential equation uniquely determine the state transition map of the dynamical system while the equation  $r(x, \dot{x}, F) = x$  determines the



readout map. Thus, the physical system is *represented* by a differential equation, a readout map, and input and output spaces.

This simple example leads us to ask a few important questions regarding representations of dynamical systems. What are common representations of dynamical systems? Linear dynamical systems? Linear, time-invariant dynamical systems?

We'll first answer these questions for the continuous-time case, in which  $\mathcal{T} = \mathbb{R}$ . In order to present a well-posed definition for system representations of continuous-time linear systems, we first need to define a special class of signals: *piecewise continuous signals*. In the next section, we'll find that piecewise continuous signals are the “right” class of input signal for continuous-time linear system representations.

**Definition 2.11 (Piecewise Continuity)** Let  $V$  be a normed vector space and  $I \subseteq \mathbb{R}$  a (possibly infinite) interval. A mapping  $u : I \rightarrow V$  is said to be piecewise continuous on  $I$  if there exists a set  $D \subseteq I$ , called the discontinuity set, for which the following hold:

1. Continuity outside  $D$ :  $u$  is continuous on  $I \setminus D$ .
2. Left and right limits: for all  $\tau \in D$ , the left and right limits  $\lim_{t \rightarrow \tau^-} u(t)$  and  $\lim_{t \rightarrow \tau^+} u(t)$  exist and are finite.
3. Finite intersections: for all  $t_0, t_1 \in \mathbb{R}$  with  $t_0 \leq t_1$ , the set  $D \cap [t_0, t_1]$  contains a finite number of points.

The set of all piecewise continuous mappings from  $I$  into  $V$  is denoted  $PC(I, V)$ .

*Remark 2.12* The above is sometimes referred to as “piecewise continuity with one-sided limits.” Here, we incorporate the one-sided limits into the definition of piecewise continuity.

*Example 2.1 (Unit Step Function)* The unit step function,  $\mathbb{1} : \mathbb{R} \rightarrow \mathbb{R}$ , defined

$$\mathbb{1}(t) = \begin{cases} 0 & t < 0 \\ 1 & t \geq 0, \end{cases} \quad (2.18)$$

is a piecewise continuous function in  $PC(\mathbb{R}, \mathbb{R})$ , with discontinuity set  $D = \{0\}$ .

**Exercise 2.5** Verify that any continuous mapping  $f : I \rightarrow V$  is piecewise continuous.

Importantly, the set of piecewise continuous functions has a natural vector space structure.

**Proposition 2.2 ( $PC(I, V)$  is a Vector Space)** Let  $I \subseteq \mathbb{R}$  a nonempty interval and  $V$  a normed vector space over  $\mathbb{K}$ . When equipped with the operations  $+$  of function addition and  $(\cdot)$  of scalar multiplication of functions,  $PC(I, V)$  forms a vector space over  $\mathbb{K}$ .

**Proof** See Problem 2.2. □

With these definitions in our toolbelt, we formulate a precise, well-posed definition for a continuous-time LTV system representation.

**Definition 2.12 (Continuous-Time LTV System Representation)** A continuous-time LTV system representation consists of the following data:

1. Input, output, and state spaces: an input space  $\mathcal{U} = PC(\mathbb{R}, \mathbb{R}^m)$ , output space  $\mathcal{Y} = PC(\mathbb{R}, \mathbb{R}^p)$ , and state space  $\Sigma = \mathbb{R}^n$ .

2. Matrix functions: matrix functions  $A(\cdot)$ ,  $B(\cdot)$ ,  $C(\cdot)$ , and  $D(\cdot)$  satisfying,

$$A(\cdot) \in PC(\mathbb{R}, \mathbb{R}^{n \times n}), B(\cdot) \in PC(\mathbb{R}, \mathbb{R}^{n \times m}), \quad (2.19)$$

$$C(\cdot) \in PC(\mathbb{R}, \mathbb{R}^{p \times n}), D(\cdot) \in PC(\mathbb{R}, \mathbb{R}^{p \times m}). \quad (2.20)$$

3. State & output equations: a differential equation and an algebraic equation,

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) \text{ (state equation)} \quad (2.21)$$

$$y(t) = C(t)x(t) + D(t)u(t) \text{ (output equation),} \quad (2.22)$$

where  $x(t) \in \mathbb{R}^n$ ,  $u(\cdot) \in \mathcal{U} = PC(\mathbb{R}, \mathbb{R}^m)$ , and  $y(\cdot) \in \mathcal{Y} = PC(\mathbb{R}, \mathbb{R}^p)$ .

We refer to the system representation by the tuple  $(A(\cdot), B(\cdot), C(\cdot), D(\cdot))$ . Such a representation is said to be a continuous-time, *state-space* system representation. The vector  $x(t)$  is referred to as the *state vector*,  $u(t)$  as the *input vector*, and  $y(t)$  as the *output vector*.

*Remark 2.13* Since the input space of a continuous-time LTV system representation is  $\mathcal{U} = PC(\mathbb{R}, \mathbb{R}^m)$ , the input-value space is  $U = \mathbb{R}^m$ . Likewise, the output-value space is  $Y = \mathbb{R}^p$ .

This definition has a lot of moving parts, so let's take a moment to summarize the key points. A continuous time, linear time-varying system is defined by a tuple  $(A(\cdot), B(\cdot), C(\cdot), D(\cdot))$  of *matrix-valued* functions, each of which is piecewise continuous. The time set of the system is  $\mathbb{R}$ , which makes it a continuous-time system. The state space of such a system is  $\mathbb{R}^n$ , while the input and output spaces are  $PC(\mathbb{R}, \mathbb{R}^m)$  and  $PC(\mathbb{R}, \mathbb{R}^p)$ . Since the system is described by a pair of a state equation (the differential equation  $\dot{x} = A(t)x + B(t)u$ ) and an output equation (the algebraic equation  $y = C(t)x + D(t)u$ ), such a representation is referred to as a *state-space* representation.

Why the emphasis on piecewise continuity? As we'll see in the next section, the piecewise continuity assumption is *essential* for the system representation to determine a unique dynamical system. Without this assumption, we aren't guaranteed to have unique solutions to the differential equation  $\dot{x} = A(t)x + B(t)u$ , which would cause us trouble in defining a state transition map.

Now that we've defined a linear, time-varying system representation, we have an enormous open question on our hands:

*Does Definition 2.12 determine a formal linear I/O dynamical system in the sense of Definition 2.8?*

In order to answer this question, one must perform a nontrivial study of the existence and uniqueness of solutions to differential equations. Nevertheless, we will find in the next section that, after performing this study, Definition 2.12 does indeed yield a valid system representation. More on this later!

Let's write down a few more important classes of system representations. Now that we've defined a continuous time, linear time-varying system representation, it's only natural to define a continuous time, linear time-invariant system representation.

**Definition 2.13 (Continuous-Time LTI System Representation)** A continuous-time LTI system representation consists of the following data:

1. Input, output, and state spaces: an input space  $\mathcal{U} = PC(\mathbb{R}, \mathbb{R}^m)$ , output space  $\mathcal{Y} = PC(\mathbb{R}, \mathbb{R}^p)$ , and state space  $\Sigma = \mathbb{R}^n$ .

2. Matrices: fixed matrices  $A \in \mathbb{R}^{n \times n}$ ,  $B \in \mathbb{R}^{n \times m}$ ,  $C \in \mathbb{R}^{p \times n}$ , and  $D \in \mathbb{R}^{p \times m}$ .  
 3. State & output equations: a differential equation and an algebraic equation,

$$\dot{x}(t) = Ax(t) + Bu(t) \text{ (state equation)} \quad (2.23)$$

$$y(t) = Cx(t) + Du(t) \text{ (output equation),} \quad (2.24)$$

where  $x(t) \in \mathbb{R}^n$ ,  $u(\cdot) \in \mathcal{U} = PC(\mathbb{R}, \mathbb{R}^m)$ , and  $y(\cdot) \in \mathcal{Y} = PC(\mathbb{R}, \mathbb{R}^p)$ .

We refer to the system representation by the tuple  $(A, B, C, D)$ .

Thus, in order to define a continuous time, linear time *invariant* system, we simply take the definition of a continuous time, linear time varying system and remove all dependence on time from the matrix functions  $(A(\cdot), B(\cdot), C(\cdot), D(\cdot))$ . Once again, we face a question regarding how this representation relates to the LTI systems we defined above.

*Does Definition 2.13 determine a formal LTI I/O dynamical system in the sense of Definition 2.9?*

Fortunately, we'll find that the answer is yes! As with the above, we'll wait until the next section to prove this.

Now, we define discrete-time analogues of Definitions 2.12 and 2.13. Since the time set of a discrete-time system is  $\mathbb{Z}$ , we can drop all of the piecewise continuity assumptions on matrix functions and signals when defining discrete-time system representations.

**Definition 2.14 (Discrete-Time LTV System Representation)** A discrete-time LTV system representation consists of the following data:

1. Input, output and state spaces: an input space  $\mathcal{U} = \{u : \mathbb{Z} \rightarrow \mathbb{R}^m\}$  and output space  $\mathcal{Y} = \{y : \mathbb{Z} \rightarrow \mathbb{R}^p\}$  of all functions from  $\mathbb{Z}$  to  $\mathbb{R}^m$  and  $\mathbb{R}^p$ , and state space  $\Sigma = \mathbb{R}^n$ .  
 2. Matrix functions: matrix-valued functions  $A[\cdot], B[\cdot], C[\cdot], D[\cdot]$ ,

$$A[\cdot] : \mathbb{Z} \rightarrow \mathbb{R}^{n \times n}, B[\cdot] : \mathbb{Z} \rightarrow \mathbb{R}^{n \times m} \quad (2.25)$$

$$C[\cdot] : \mathbb{Z} \rightarrow \mathbb{R}^{p \times n}, D[\cdot] : \mathbb{Z} \rightarrow \mathbb{R}^{p \times m}. \quad (2.26)$$

3. State & output equations: a recurrence relation and an algebraic equation,

$$x[k+1] = A[k]x[k] + B[k]u[k] \text{ (state equation)} \quad (2.27)$$

$$y[k] = C[k]x[k] + D[k]u[k] \text{ (output equation),} \quad (2.28)$$

where  $x[k] \in \mathbb{R}^n$ ,  $u[\cdot] \in \mathcal{U} = \{u : \mathbb{Z} \rightarrow \mathbb{R}^m\}$ , and  $y[\cdot] \in \mathcal{Y} = \{y : \mathbb{Z} \rightarrow \mathbb{R}^p\}$ .

We refer to the system representation by the tuple  $(A[\cdot], B[\cdot], C[\cdot], D[\cdot])$ . Such a system is said to be a discrete time, *state-space* system representation. The vector  $x[k]$  is referred to as the *state vector*,  $u[k]$  as the *input vector*, and  $y[k]$  as the *output vector*.

*Remark 2.14* Note that there are several popular ways of writing the state and output equations of a discrete-time system. Above, we've used square brackets to describe the time,  $k \in \mathbb{Z}$ . Other popular notation includes,

$$x_{k+1} = A_k x_k + B_k u_k \quad x(k+1) = A(k)x(k) + B(k)u(k) \quad (2.29)$$

$$y_k = C_k x_k + D_k u_k \quad y(k) = C(k)x(k) + D(k)u(k). \quad (2.30)$$

*Remark 2.15* In Definition 2.14, we’ve defined inputs, outputs, and matrix-valued functions to be functions from  $\mathbb{Z}$  into each of their respective spaces. This means that we can view the inputs, outputs, and matrix functions as *sequences*.

**Definition 2.15 (Discrete-Time LTI System Representation)** A discrete-time LTI system representation consists of the following data:

1. Input, output, and state spaces: an input space  $\mathcal{U} = \{u : \mathbb{Z} \rightarrow \mathbb{R}^m\}$ , output space  $\mathcal{Y} = \{y : \mathbb{Z} \rightarrow \mathbb{R}^p\}$ , and state space  $\Sigma = \mathbb{R}^n$ .
2. Matrices: fixed matrices  $A \in \mathbb{R}^{n \times n}$ ,  $B \in \mathbb{R}^{n \times m}$ ,  $C \in \mathbb{R}^{p \times n}$ , and  $D \in \mathbb{R}^{p \times m}$ .
3. State & output equations: a recurrence relation and an algebraic equation,

$$x[k+1] = Ax[k] + Bu[k] \text{ (state equation)} \quad (2.31)$$

$$y[k] = Cx[k] + Du[k] \text{ (output equation),} \quad (2.32)$$

where  $x[k] \in \mathbb{R}^n$ ,  $u[\cdot] \in \mathcal{U} = \{u : \mathbb{Z} \rightarrow \mathbb{R}^m\}$ , and  $y[\cdot] \in \mathcal{Y} = \{y : \mathbb{Z} \rightarrow \mathbb{R}^p\}$ .

We refer to the system representation by the tuple  $(A, B, C, D)$ .

The fact that we don’t have to worry about regularity conditions on our signals (as we did with piecewise-continuity in continuous-time) hints that the analysis of discrete-time systems might be easier than the analogous analysis of continuous-time systems. This is in fact the case in a number of scenarios.

To wrap this section up, we define SISO and MIMO systems. Frequently, we’ll distinguish between systems that only have a single input and output (which are generally easier to analyze) and systems that have multiple inputs and outputs.

**Definition 2.16 (SISO/MIMO System Representations)** Consider a continuous or discrete-time system representation in which

$$\mathcal{U} = PC(\mathbb{R}, \mathbb{R}^m) \text{ or } \mathcal{U} = \{u : \mathbb{Z} \rightarrow \mathbb{R}^m\} \quad (2.33)$$

$$\mathcal{Y} = PC(\mathbb{R}, \mathbb{R}^p) \text{ or } \mathcal{Y} = \{y : \mathbb{Z} \rightarrow \mathbb{R}^p\}. \quad (2.34)$$

If  $m = p = 1$ , the system is said to be single-input, single-output (SISO). If  $m, p \geq 1$ , the system is said to be multi-input, multi-output (MIMO).

*Remark 2.16* Note that we don’t take the definition of MIMO to be strictly greater than 1—this way, we can consider MIMO to be a straightforward generalization of SISO.

Based on this definition, SISO systems appear to be scalar from the input-output perspective: we put in a scalar as an input and get another scalar as an output. It’s important to note that the SISO/MIMO distinction has *nothing* to do with the dimension of the state space! One can have a SISO system with an arbitrarily high-dimension state space, provided the input and output-value spaces are both one-dimensional.

### 2.1.3 Further Reading

This section was mainly influenced by [36], [27], and [8]. Our development of abstract dynamical systems most closely follows that of Chapter 5 of [8]. For a more in-depth look at abstract dynamical systems, we refer the reader to Chapter 2 of [36]. The example of a shower control system, used all the way at the start of the section, is from [30].

### 2.1.4 Problems

**Problem 2.1 (Causal & Noncausal Maps [2])** As we mentioned in the section above, one can represent the input/output relationship of a system for a *fixed* initial time and state directly with a function  $H : \mathcal{T} \times \mathcal{U} \rightarrow \mathcal{Y}$ . That is, one has  $y(t) = H(t, u(\cdot))$  for any time  $t$  and admissible input  $u(\cdot)$ . In this problem, we'll determine definitions for causality, linearity, and time-invariance of an arbitrary map  $H : \mathcal{T} \times \mathcal{U} \rightarrow \mathcal{Y}$ .

1. Given an arbitrary map  $H : \mathcal{T} \times \mathcal{U} \rightarrow \mathcal{Y}$ , formulate a definition of *time-invariance* for  $H$ . Formulate a definition of *causality*. Formulate a definition of *linearity*. *Hint: for causality, think about the restriction of a signal to a certain time interval.*
2. Let's put our definitions to the test. In each of the following cases, determine whether the system is causal/time-invariant/linear. Use your best judgment to identify the input and output spaces in each case.
  - a. Consider a discrete-time system with I/O description  $y[k] = c_1 u[k+1] + c_2$ , where  $c_1, c_2 \in \mathbb{R}$ . Is this system causal? Is it time-invariant? Is it linear?
  - b. Consider a continuous-time system with I/O description  $y(t) = u(t - \tau)$ , where  $\tau \in \mathbb{R}$  is fixed and positive. Is this system causal? Is it time invariant? Is it linear?
  - c. Consider a continuous-time system with I/O description,

$$y(t) = \begin{cases} u(t) & t \leq \tau \\ 0 & t > \tau. \end{cases} \quad (2.35)$$

Is this system causal? Is it time-invariant? Is it linear?

- d. Consider a continuous-time system with I/O description,

$$y(t) = \min\{u_1(t), u_2(t)\}, \quad (2.36)$$

where  $u(t) = [u_1(t); u_2(t)]^\top$  is the system input. Is this system causal? Is it time-invariant? Is it linear?

**Problem 2.2 (Properties of Piecewise-Continuous Functions)** In the section above, we introduced the class of piecewise-continuous functions. In this problem, we'll prove some basic properties of this function class.

1. Show that  $PC(\mathbb{R}, \mathbb{R}^n)$  forms a vector space over  $\mathbb{R}$  under the operations of function addition and scalar multiplication.
2. Let  $I, K \subseteq \mathbb{R}$  be compact intervals. Show that any  $f \in PC(I, \mathbb{R})$  must be bounded above on  $I \cap K$ ,

$$\sup_{t \in I \cap K} f(t) < \infty. \quad (2.37)$$

3. Let  $I \subseteq \mathbb{R}$  be a compact interval and  $\|\cdot\|$  be an arbitrary norm on  $\mathbb{R}^n$ . Show that the supremum norm,

$$\|f\|_\infty = \sup_{t \in I} \|f(t)\|, \quad (2.38)$$

is finite for all  $f \in PC(I, \mathbb{R}^n)$ . Then, prove that  $\|\cdot\|_\infty$  makes  $PC(I, \mathbb{R}^n)$  into a normed vector space.

4. Is  $PC(I, \mathbb{R}^n)$  a Banach space with respect to the supremum norm  $\|\cdot\|_\infty$ ,  $\|f\|_\infty = \sup_{t \in \mathbb{R}} \|f(t)\|$ ? Provide a proof or a counterexample.

## 2.2 Solutions of Linear, Time-Varying Systems

Now that we've introduced a set of state space representations of linear systems, we must show that these representations *are* in fact representations in the formal sense. Recall that in the previous section, we posed two questions:

*Does an LTV system representation determine a linear I/O system?*

*Does an LTI system representation determine an LTI I/O system?*

Further, we promised that in this section, we would provide a *precise* answer to both of these queries. Now that we're here, we need to make good on this promise! In this section, we establish answers to these questions by studying the existence, uniqueness, and structure of solutions to linear ordinary differential equations and recurrence relations. We'll then apply the results of this study to answer the two questions above.

In order to answer these questions, we'll split up into the continuous and discrete-time cases. In order to study continuous-time linear systems, we must study linear ordinary differential equations, while to study discrete-time linear systems, we must study linear recurrence relations. Along the way, we'll draw connections between the techniques used to study the two. Let's begin!

### 2.2.1 Solutions of Continuous-Time Linear Systems

We'll begin by laying out a brief plan of attack for our study of continuous-time linear system representations. Recall from the previous section that a continuous-time LTV system representation is specified by a tuple  $(A(\cdot), B(\cdot), C(\cdot), D(\cdot))$  of piecewise continuous matrix-valued functions. These functions define two equations,

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) \text{ (state equation)} \quad (2.39)$$

$$y(t) = C(t)x(t) + D(t)u(t) \text{ (output equation),} \quad (2.40)$$

which govern how the state  $x(t)$  and the output  $y(t)$  change over time as input signals  $u : \mathbb{R} \rightarrow \mathbb{R}^m$  are applied to the system.

In order to determine if the LTV system representation yields a valid linear input/output dynamical system in the sense of Definition 2.8, we must tick a couple of boxes. To verify that the LTV system representation yields a linear I/O system, we must compute the I/O map,  $\rho$ , associated to the representation, and verify that it satisfies the linearity conditions proposed in the previous section. In order to compute  $\rho$ , however, we require the state transition map,  $\varphi$ , of the representation.

Thus, we begin by studying the state transition map. The state transition map associated to the LTV system representation maps from times  $t_0, t_1 \in \mathbb{R}$  with  $t_0 \leq t_1$ , initial state  $x_0 \in \mathbb{R}^n$ , and input signal  $u \in PC(\mathbb{R}, \mathbb{R}^m)$  to the solution of the differential equation,

$$\dot{x}(t) = A(t)x(t) + B(t)u(t), \quad (2.41)$$

at time  $t_1$ , with initial condition  $x(t_0) = x_0$ . Therefore, in order to have a state transition map  $\varphi(t_1, t_0, x_0, u(\cdot))$  that is well-defined on the input space  $\mathcal{U} = PC(\mathbb{R}, \mathbb{R}^m)$ , there are a couple of things we need:

1. Existence: for all inputs  $u \in PC(\mathbb{R}, \mathbb{R}^m)$ , initial conditions  $x_0 \in \mathbb{R}^n$ , and times  $t_0 \in \mathbb{R}$ , a solution to the initial value problem,

$$\dot{x}(t) = A(t)x(t) + B(t)u(t), \quad x(t_0) = x_0, \quad (2.42)$$

must exist. This ensures that we will always be able to compute  $\varphi$  on the time set, state space, and input space of the representation.

2. Uniqueness: for all inputs  $u \in PC(\mathbb{R}, \mathbb{R}^m)$ , initial conditions  $x_0 \in \mathbb{R}^n$ , and times  $t_0 \in \mathbb{R}$ , the solution to the initial value problem,

$$\dot{x}(t) = A(t)x(t) + B(t)u(t), \quad x(t_0) = x_0, \quad (2.43)$$

must be *unique*. If we want the state transition map to be well-defined, we can't have two or more solutions to the initial value problem!

Thus, in order to solve the problem of verifying the LTV system representation yields a linear I/O system, we must first establish the existence and uniqueness of solutions to the initial value problem  $\dot{x} = A(t)x(t) + B(t)u(t)$ ,  $x(t_0) = x_0$ . We'll break this down into the following multi-step process:

1. Define Solutions: first, we'll formulate a precise definition for a solution to a time-varying initial value problem with piecewise continuous data.
2. Matrix IVP: next, we'll argue that it will be insightful to study solutions to a simpler yet more fundamental initial value problem, the *matrix* initial value problem

$$\dot{X}(t) = A(t)X(t), \quad X(t_0) = I. \quad (2.44)$$

We will prove that this problem has a unique solution and will study its basic structure.

3. State Transition Matrix: after showing that a unique solution to the IVP  $\dot{X}(t) = A(t)X(t)$ ,  $X(t_0) = I$  exists, we'll *abstract away* details of the solution into a special operator called the state transition matrix. Then, we'll establish a few important properties of the state transition matrix.
4. LTV-IVP: finally, we'll show that we can use the state transition matrix to study solutions to the general linear time-varying, initial value problem. Here, we'll complete the problem of proving existence and uniqueness of solutions.

### 2.2.1.1 Defining Solutions to IVPs

Let's tackle the first step of the process we outlined above. What does it mean to solve a time-varying initial value problem with piecewise continuous data? Although it might *seem* like all we need is to find a differentiable function which satisfies  $\dot{x}(t) = A(t)x(t) + B(t)u(t)$  for all  $t \in \mathbb{R}$  and  $x(t_0) = x_0$ , the reality is somewhat more complex! To illustrate what goes wrong with this "naive" definition of a solution, consider the case of a simple scalar, time-varying initial value problem,

$$\dot{x}(t) = b(t), \quad x(0) = 0. \quad (2.45)$$

Suppose  $b \in PC(\mathbb{R}, \mathbb{R})$  is the *step function*, the function which is identically zero before  $t = 0$  and identically one at and after  $t = 0$ ,

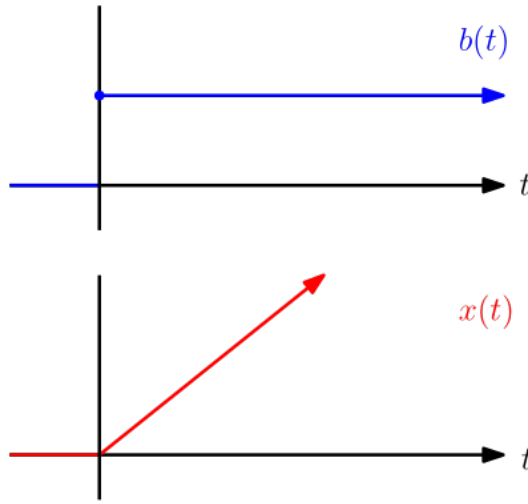


$$b(t) = \begin{cases} 1, & t \geq 0 \\ 0, & t < 0. \end{cases} \quad (2.46)$$

Using our basic knowledge of ODEs, we know that the solution to this initial value problem *should* be the ramp function,

$$x(t) = \begin{cases} t & t \geq 0 \\ 0 & t < 0. \end{cases} \quad (2.47)$$

However, this function is clearly *not* differentiable at the point  $t = 0$ —right where  $b$  makes



**Fig. 2.4** The solution to the initial value problem  $\dot{x}(t) = b(t)$ , with  $b$  the step function, *should* be the ramp function. However, the ramp function is *not* differentiable at  $t = 0$ ! Thus, our definition for a solution an IVP must account for points of non-differentiability.

its jump, we find that the proposed solution of the initial value problem has a sharp “corner.” As a result of this, we find that requiring a solution to this IVP to be a *differentiable* function  $x : \mathbb{R} \rightarrow \mathbb{R}$  satisfying  $\dot{x}(t) = b(t) \forall t$ ,  $x(0) = 0$ , is *too strong*.

This example is a particular instance of a general fact from analysis: if a function has a jump discontinuity, it *cannot* be the derivative of another function (see Problem 2.4 for the formal details of this argument). We conclude that, in order to make a *well-posed* definition for a solution to an initial value problem with piecewise continuous data, we must explicitly account for the points of discontinuity. This leads us to the following, formal definition of a solution.

**Definition 2.17 (Solution to LTV-IVP)** Consider the piecewise-continuous maps  $A(\cdot) \in PC(\mathbb{R}, \mathbb{R}^{n \times n})$ ,  $B(\cdot) \in PC(\mathbb{R}, \mathbb{R}^{n \times m})$ , and  $u(\cdot) \in PC(\mathbb{R}, \mathbb{R}^m)$ , which have a shared discontinuity set  $D \subseteq \mathbb{R}$ . For  $x_0 \in \mathbb{R}^n$  and  $t_0 \in \mathbb{R}$ , a solution to the initial value problem,

$$\dot{x}(t) = A(t)x(t) + B(t)u(t), \quad x(t_0) = x_0, \quad (2.48)$$

is a continuous map  $x : \mathbb{R} \rightarrow \mathbb{R}^n$ , satisfying the conditions:

1. Initial condition:  $x(t_0) = x_0$ .
2. Derivative: For all  $t \in \mathbb{R} \setminus D$ ,  $\frac{d}{dt}x(t) = A(t)x(t) + B(t)u(t)$ .

*Remark 2.17* Here, we take solutions to be defined for *all*  $t \in \mathbb{R}$ . We'll later see that this is justified for the case of LTV systems with piecewise continuous data. However, this requirement should be relaxed for *nonlinear* initial value problems.

*Remark 2.18* Here, we assume that each map has the same discontinuity set - this assumption is made without loss of generality, since one can always take the union of the discontinuity sets of  $A(\cdot), B(\cdot), u(\cdot)$  if they do not initially coincide.

This formal definition of a solution to an initial value problem *relaxes* our “intuitive” definition of a solution. It tells us that we only need to check the derivative condition at times when all data defining the initial value problem is continuous. Since carrying around the discontinuity set  $D$  can get a little cumbersome, we state an *equivalent* definition of a solution to an initial value problem which doesn't require the use of  $D$ . We state this definition in the form of a proposition.

**Proposition 2.3 (Integral Solution of IVPs)** *Consider the piecewise continuous maps  $A(\cdot) \in PC(\mathbb{R}, \mathbb{R}^{n \times n})$ ,  $B(\cdot) \in PC(\mathbb{R}, \mathbb{R}^{n \times m})$ , and  $u \in PC(\mathbb{R}, \mathbb{R}^m)$ . A continuous function  $x : \mathbb{R} \rightarrow \mathbb{R}^n$  is a solution to the initial value problem,*

$$\dot{x}(t) = A(t)x(t) + B(t)u(t), \quad x(t_0) = x_0, \quad (2.49)$$

*if and only if, for all  $t \in \mathbb{R}$ , it satisfies*

$$x(t) = x_0 + \int_{t_0}^t A(\tau)x(\tau) + B(\tau)u(\tau)d\tau. \quad (2.50)$$

The proof of this proposition is essentially a straightforward application of the fundamental theorem of calculus, which we now recall.

**Theorem 2.1 (Fundamental Theorem of Calculus)** . *Let  $f : \mathbb{R} \rightarrow \mathbb{R}^n$  be a Riemann-integrable function.<sup>2</sup> Then, the following results hold:*

1. *Fix a number  $t_0 \in \mathbb{R}$  and define  $F(t) = \int_{t_0}^t f(\tau)d\tau$ . Then,  $F$  is continuous. Further, if  $f$  is continuous at  $t$ , then  $F$  is differentiable at  $t$ , with  $F'(t) = f(t)$ .*
2. *If  $F : \mathbb{R} \rightarrow \mathbb{R}^n$  is a Riemann-integrable function satisfying  $F'(t) = f(t)$  for all but a finite number of points in an interval  $[t_0, t_1] \subseteq \mathbb{R}$ , then  $\int_{t_0}^{t_1} f(\tau)d\tau = F(t_1) - F(t_0)$ .*

**Proof** See [1] for details. □

Now, we return to the proof of Proposition 2.3.

**Proof (Of Proposition 2.3)** Suppose  $x : \mathbb{R} \rightarrow \mathbb{R}^n$  is a solution to the initial value problem in the sense of Definition 2.17. Then, for  $D$  the shared discontinuity set of  $A(\cdot), B(\cdot), u(\cdot)$ , one has that  $\dot{x}(t) = A(t)x(t) + B(t)u(t)$ , for all  $t \in \mathbb{R} \setminus D$ . We aim to show that  $x$  satisfies the integral condition proposed above.

Fix a time  $t \in \mathbb{R}$ , assuming  $t \geq t_0$  (the proof is identical for  $t < t_0$ ). By definition of a piecewise continuous function, it follows that  $[t_0, t] \cap D$  only contains a finite number of

<sup>2</sup> We say that a function  $f : \mathbb{R} \rightarrow \mathbb{R}^n$  is Riemann-integrable if its Riemann integral  $\int_a^b f(t)dt$  is defined for all (finite)  $a, b \in \mathbb{R}$ . One may show that all piecewise continuous functions are Riemann-integrable.

points. Thus, between  $t_0$  and  $t_1$ ,  $\dot{x}(t) = A(t)x(t) + B(t)u(t)$  at all but a finite number of points. By the fundamental theorem of calculus, it then follows that

$$x(t) = x(t_0) + \int_{t_0}^t A(\tau)x(\tau) + B(\tau)u(\tau)d\tau = x_0 + \int_{t_0}^t A(\tau)x(\tau) + B(\tau)u(\tau)d\tau. \quad (2.51)$$

This completes the first direction of the proof. Now, we proceed in the other direction. Suppose  $x(t) = x_0 + \int_{t_0}^t A(\tau)x(\tau) + B(\tau)u(\tau)d\tau$  for all  $t \geq t_0$ . Taking  $t = t_0$ , one gets that  $x(t_0) = x_0$ , yielding the initial condition constraint. Now, we focus on differentiability. We know that  $A(t)x(t) + B(t)u(t)$  must be continuous at all  $t \in \mathbb{R} \setminus D$ , since  $x$  is continuous by the fundamental theorem and  $A(\cdot), B(\cdot), u(\cdot)$  are continuous on  $\mathbb{R} \setminus D$ . By the fundamental theorem,  $x$  is differentiable on  $\mathbb{R} \setminus D$  and satisfies  $\dot{x}(t) = A(t)x(t) + B(t)u(t)$  for all  $t \in \mathbb{R} \setminus D$ . We conclude that  $x$  is a solution to the initial value problem.  $\square$

### 2.2.1.2 A Matrix Initial Value Problem & The Peano-Baker Series

In the previous section, we wrote a formal definition for a solution to an initial value problem and formulated an equivalent definition of a solution using integration. In this subsection, we will examine a special, simple initial value problem that will provide almost *complete* insight into the existence and uniqueness problem we're aiming to solve.

Since the initial value problem  $\dot{x}(t) = A(t)x(t) + B(t)u(t)$ ,  $x(t_0) = x_0$  seems quite complex to analyze, having a number of moving parts, it might be beneficial to start with a simpler problem. Let's drop the input term, and study solutions to the initial value problem,

$$\dot{x}(t) = A(t)x(t), \quad x(t_0) = x_0. \quad (2.52)$$

Can we simplify this problem even further? Using the integral definition of a solution to an initial value problem, we make the following observation.

**Lemma 2.1 (Matrix IVP/Vector IVP)** *Consider the initial value problem,*

$$\dot{x}(t) = A(t)x(t), \quad x(t_0) = x_0. \quad (2.53)$$

*If  $X : \mathbb{R} \rightarrow \mathbb{R}^{n \times n}$  solves the matrix initial value problem,*

$$\dot{X}(t) = A(t)X(t), \quad X(t_0) = I, \quad (2.54)$$

*then  $x(t) = X(t)x_0$  solves the vector initial value problem  $\dot{x}(t) = A(t)x(t)$ ,  $x(t_0) = x_0$ .*

**Remark 2.19** In this lemma, we use a solution to a *matrix* initial value problem. Solutions to such initial value problems are defined identically to vector initial value problems. In fact, we can write down a vector initial value problem corresponding to any given matrix initial value problem by stretching the matrix out into a vector. Because of this equivalence, the integral definition of a solution to an IVP still holds in the matrix case. Try writing down a formal definition of a solution to a matrix IVP to check your understanding!

**Proof** Suppose  $X(t)$  solves the matrix IVP defined in the statement of the lemma. Then,

$$X(t) = I + \int_{t_0}^t A(\tau)X(\tau)d\tau. \quad (2.55)$$

Multiplying by  $x_0$ , one has,

$$X(t)x_0 = x_0 + \int_{t_0}^t A(\tau)X(\tau)x_0 d\tau, \quad (2.56)$$

which implies  $x(t) = X(t)x_0$  is a solution to the IVP  $\dot{x}(t) = A(t)x(t)$ ,  $x(t_0) = x_0$ .  $\square$

This lemma yields a great deal of insight into the structure of solutions to  $\dot{x}(t) = A(t)x(t)$ ,  $x(t_0) = x_0$ . In particular, it tells us that there exist solutions to the initial value problem that are *linear* in the initial condition! Further, these solutions are *entirely* determined by solutions to the *matrix* IVP  $\dot{X}(t) = A(t)X(t)$ ,  $X(t_0) = I$ . Thus, in order to understand solutions to the vector initial value problem  $\dot{x}(t) = A(t)x(t)$ ,  $x(t_0) = x_0$ , we will study solutions to the associated *matrix* initial value problem,  $\dot{X}(t) = A(t)X(t)$ ,  $X(t_0) = I$ .

What do we know about solutions to this matrix IVP? Do solutions exist? If so, what form do they take? A solution to the matrix IVP must satisfy the integral equation,

$$X(t) = I + \int_{t_0}^t A(\tau)X(\tau) d\tau. \quad (2.57)$$

We notice that  $X(\cdot)$  appears both on the left and right hand sides of the expression. Let's try re-plugging in the integral form of the solution into the  $X(\tau)$  on the right hand side. This gives,

$$X(t) = I + \int_{t_0}^t A(\tau) \left[ I + \int_{t_0}^{\tau} A(\tau')X(\tau') d\tau' \right] d\tau \quad (2.58)$$

$$= I + \left[ \int_{t_0}^t A(\tau) d\tau \right] + \int_{t_0}^t A(\tau) \left[ \int_{t_0}^{\tau} A(\tau')X(\tau') d\tau' \right] d\tau. \quad (2.59)$$

Interestingly, what we get inside the larger integral is the *same* expression that we originally substituted into. Thus, if we substitute again—this time for  $X(\tau')$ —we would find the same pattern! Indefinitely performing this substitution leads to the following definition.

**Definition 2.18 (Peano-Baker Series)** Let  $A(\cdot) \in PC(\mathbb{R}, \mathbb{R}^{n \times n})$ . The Peano-Baker series with respect to  $A(\cdot)$  is the infinite matrix series,

$$\Phi(t, t_0) = \sum_{k=0}^{\infty} S_k(t, t_0), \quad (2.60)$$

whose summands are defined by the recurrence,

$$S_0 = I, \quad S_{k+1}(t, t_0) = \int_{t_0}^t A(\tau)S_k(\tau, t_0) d\tau. \quad (2.61)$$

In order to confirm that this definition is well-posed, we must, we must confirm that the Peano-Baker series actually converges. A useful tool for certifying the uniform convergence<sup>3</sup> of a series of functions is the *Weierstrass M-Test*.

<sup>3</sup> Recall from Chapter 1 that a sequence of functions  $f_n : I \subseteq \mathbb{R} \rightarrow V$  (for  $(V, \|\cdot\|)$  a normed vector space) converges *uniformly* if it converges with respect to the sup norm,  $\|f\|_{\infty} = \sup_{t \in I} \|f(t)\|$ .

**Theorem 2.2 (Weierstrass M-Test)** *Let  $(V, \|\cdot\|)$  be a finite dimensional, normed vector space. Let  $\{f_n\}$  be a collection of mappings  $f_n : A \rightarrow V$  on a set  $A$ . Let  $\{M_n\} \subseteq \mathbb{R}$  be a sequence for which  $\sup_{t \in A} \|f_n(t)\| \leq M_n$ . If  $\sum_{n=1}^{\infty} M_n$  converges, then  $\sum_{n=1}^{\infty} f_n(t)$  converges uniformly on  $A$ .*

**Proof** See [1] for the details.  $\square$

In order to apply the Weierstrass M-test to certify the convergence of the Peano-Baker series, we require the following lemma.

**Lemma 2.2 (Suprema of Piecewise Continuous Functions)** *Consider a piecewise continuous function  $f \in PC(I, \mathbb{R})$ , where  $I \subseteq \mathbb{R}$ . For any  $K \subseteq I$  which is compact in  $\mathbb{R}$ ,*

$$\sup_{t \in K \cap I} f(t) < \infty \quad (2.62)$$

**Proof** See Problem 2.2.  $\square$

With these results in mind, we study the convergence of the Peano-Baker series.

**Proposition 2.4 (Convergence of Peano-Baker Series)** *Let  $A(\cdot) \in PC(\mathbb{R}, \mathbb{R}^{n \times n})$ . For any  $t' \in \mathbb{R}_{\geq 0}$ , the Peano-Baker series  $\Phi(\cdot, \cdot)$  defined by  $A(\cdot)$  converges uniformly on  $[-t', t']^2$ .*

*Remark 2.20* The notation  $[-t, t']^2$  refers to the Cartesian product  $[-t', t'] \times [-t', t']$ .

**Proof** Our proof follows the Weierstrass M-test. Fix an interval  $[-t', t'] \subseteq \mathbb{R}$  and an initial time  $t_0 \in [-t', t']$ . Recall that the Peano-Baker series at time  $t \in [-t', t']$  is defined,

$$\Phi(t, t_0) = \sum_{k=0}^{\infty} S_k(t, t_0), \quad S_0 = I, \quad S_k(t, t_0) = \int_{t_0}^t A(\tau) S_{k-1}(\tau, t_0) d\tau, \quad (2.63)$$

In order to prove that this series converges uniformly using the Weierstrass M-test, we'll exhibit a uniform bound on each  $S_k(\cdot, \cdot)$ . First, we'll show that the bound,

$$\|S_k(t, t_0)\| \leq \frac{1}{k!} \left( \sup_{t \in [-t', t']} \|A(t)\| \right)^k |t - t_0|^k, \quad \forall t \in [-t', t'], \quad (2.64)$$

must hold. Notice that  $\sup_{t \in [-t', t']} \|A(t)\|$  is finite by Lemma 2.2. Note that—although it seems like we're pulling this bound out of thin air—this is actually something one can discover by playing around with the first few terms of the series. You're encouraged to try this if you're not convinced!

Let's prove that the bound holds by induction on  $k$ . The base case,  $k = 0$ , is trivial. We get  $\|S_0\| = \|I\| = 1$ , which matches the proposed bound exactly. Let  $k \geq 1$ , and assume for induction that the proposed bound holds. Now, we bound  $S_{k+1}$  for arbitrary  $t, t_0$ —keep in mind, we may have  $t \geq t_0$  or  $t < t_0$ . We have,

$$\|S_{k+1}(t, t_0)\| = \left\| \int_{t_0}^t A(\tau) S_k(\tau, t_0) d\tau \right\| \quad (2.65)$$

$$\leq \left| \int_{t_0}^t \|A(\tau)\| \|S_k(\tau, t_0)\| d\tau \right| \quad (2.66)$$

$$\leq \left| \int_{t_0}^t \sup_{\tau \in [-t', t']} \|A(\tau)\| \cdot \frac{1}{k!} \left( \sup_{\tau \in [-t', t']} \|A(\tau)\| \right)^k |\tau - t_0|^k d\tau \right|. \quad (2.67)$$

Now, we split into two cases. First, assume  $t \geq t_0$ . In this case, we have that the above is bounded,

$$\|S_{k+1}(t, t_0)\| \leq \frac{1}{k!} \left( \sup_{t \in [-t', t']} \|A(t)\| \right)^{k+1} \left| \int_{t_0}^t |\tau - t_0|^k d\tau \right| \quad (2.68)$$

$$= \frac{1}{k!} \left( \sup_{t \in [-t', t']} \|A(t)\| \right)^{k+1} \frac{1}{k+1} |t - t_0|^{k+1} \quad (2.69)$$

$$= \frac{1}{(k+1)!} \left( \sup_{t \in [-t', t']} \|A(t)\| \right)^{k+1} |t - t_0|^{k+1}, \quad \forall t \in [t_0, t_1]. \quad (2.70)$$

Thus, the proposed bound holds for  $t \geq t_0$ . For  $t < t_0$ , the same procedure is followed—simply flip  $t_0$  and  $t$  in the integral and perform the same bounds. So, by induction on  $k$ , we conclude that the proposed bound holds for all  $k \in \mathbb{N}$ . We can then bound  $\|S_k(t, t_0)\|$  uniformly in  $t, t_0$  on  $[-t', t']$  by,

$$\sup_{t, t_0 \in [-t', t']} \|S_k(t, t_0)\| \leq \frac{1}{k!} \left( \sup_{t \in [-t, t]} \|A(t)\| \right)^k (2t')^k. \quad (2.71)$$

Now, we're ready to apply the Weierstrass M-test. Define  $M_k$  as the right hand side of the inequality above. Does  $\sum_{k=0}^{\infty} M_k$  converge? We recognize the sum as the Taylor series definition of the *exponential*! Thus, we have,

$$\sum_{k=0}^{\infty} \frac{1}{k!} \left( \sup_{t \in [-t', t']} \|A(t)\| \right)^k (2t')^k = \exp \left( \sup_{t \in [-t', t']} \|A(t)\| (2t') \right) < \infty. \quad (2.72)$$

By the Weierstrass M-test, we conclude that the Peano-Baker series  $\Phi(\cdot, \cdot)$  converges uniformly on any compact interval  $[-t', t']^2$ .  $\square$

Let's summarize what we've done so far. In analyzing solutions to the matrix IVP,  $\dot{X}(t) = A(t)X(t)$ ,  $X(t_0) = I$ , we discovered a recurrent pattern that led us to the Peano-Baker series. Then, we proved that the Peano-Baker series converges uniformly on any compact interval. Now, we ask the question—does it converge to the solution of the matrix IVP? The following theorem provides an answer.

**Proposition 2.5 (Peano-Baker Series Solves the Matrix IVP)** *Let  $A \in PC(\mathbb{R}, \mathbb{R}^{n \times n})$ . Consider the matrix initial value problem  $\dot{X}(t) = A(t)X(t)$ ,  $X(t_0) = I$ . The Peano-Baker series  $\Phi(t, t_0)$  defined by  $A(\cdot)$  solves the matrix initial value problem.*

**Proof** We can prove this either via direct differentiation or the integral method. Let's proceed via differentiation. Let  $D \subseteq \mathbb{R}$  be the discontinuity set of  $A$ . By definition, one has that for all  $t \in \mathbb{R} \setminus D$ , each term of the Peano-Baker series is differentiable in its first argument. Thus, for any time  $t \in \mathbb{R}$ ,

$$\frac{d}{dt} \sum_{k=0}^n S_k(t, t_0) = \sum_{k=0}^n \frac{d}{dt} S_k(t, t_0) \quad (2.73)$$

$$= \sum_{k=1}^n \frac{d}{dt} \int_{t_0}^t A(\tau) S_{k-1}(\tau, t_0) d\tau \quad (2.74)$$

$$= A(t) \sum_{k=1}^n S_{k-1}(t, t_0) \quad (2.75)$$

$$= A(t) \sum_{k=0}^{n-1} S_k(t, t_0). \quad (2.76)$$

Now, fix an interval  $[-t_1, t_1] \subseteq \mathbb{R}$ , for which  $t_0 \in [-t_1, t_1]$ . We know that the Peano-Baker series converges uniformly in  $t$  on this interval. Additionally, the chain of equalities above implies that  $\sum_{k=0}^n \frac{d}{dt} S_k(t, t_0)$  converges uniformly in  $t$  on  $[-t_1, t_1]$ . Thus, by Theorem 1.3 on uniform convergence and differentiation, it follows that  $\frac{d}{dt} \lim_{n \rightarrow \infty} \sum_{k=0}^n S_k(t, t_0) = \lim_{n \rightarrow \infty} \sum_{k=0}^n \frac{d}{dt} S_k(t, t_0)$ , for all  $t \in (-t_1, t_1) \cap (\mathbb{R} \setminus D)$ . This means,

$$\frac{d}{dt} \sum_{k=0}^{\infty} S_k(t, t_0) = A(t) \lim_{n \rightarrow \infty} \sum_{k=0}^{n-1} S_k(t, t_0) = A(t) \sum_{k=0}^{\infty} S_k(t, t_0), \quad (2.77)$$

for all  $t \in (-t_1, t_1) \cap (\mathbb{R} \setminus D)$ . Thus, we have that,

$$\frac{d}{dt} \Phi(t, t_0) = A(t) \Phi(t, t_0), \quad \forall t \in (-t_1, t_1) \cap (\mathbb{R} \setminus D). \quad (2.78)$$

Since the interval  $[-t_1, t_1]$  can be made arbitrarily large, we conclude that the Peano-Baker series satisfies the derivative property of the initial value problem for all  $t \in \mathbb{R} \setminus D$ . Further, we have that  $\Phi(t_0, t_0) = I$  by definition. We conclude that the Peano-Baker series solves the matrix initial value problem.  $\square$

We've now established that the Peano-Baker series is *a* solution to the matrix initial value problem—is it the *only* solution? The following inequality helps us answer this question.

**Lemma 2.3 (Gronwall Inequality)** *Let  $y, k \in PC(\mathbb{R}, \mathbb{R}_{\geq 0})$  and  $c \in \mathbb{R}_{\geq 0}$ , and  $t_0 \in \mathbb{R}$ . If for all  $t \in \mathbb{R}$ ,  $y$  satisfies,*

$$y(t) \leq c + \left| \int_{t_0}^t k(\tau) y(\tau) d\tau \right|, \quad (2.79)$$

*then for all  $t \in \mathbb{R}$ ,*

$$y(t) \leq c \exp \left| \int_{t_0}^t k(\tau) d\tau \right|. \quad (2.80)$$

**Proof** See Problem 2.5.  $\square$

Using the Gronwall inequality, we prove that solutions to the matrix IVP are *unique*.

**Theorem 2.3 (Existence & Uniqueness of Solutions to Matrix IVP)** Let  $A(\cdot) \in PC(\mathbb{R}, \mathbb{R}^{n \times n})$ . The Peano-Baker series,  $\Phi(t, t_0)$ , is the unique solution to the matrix initial value problem  $\dot{X}(t) = A(t)X(t)$ ,  $X(t_0) = I$ .

**Proof** By Proposition 2.5, we already know  $\Phi(t, t_0)$  is a solution the IVP. Now, we show it is the unique solution. Suppose  $X : \mathbb{R} \rightarrow \mathbb{R}^{n \times n}$  is another solution. Then, for any  $t, t_0$ , both  $X$  and  $\Phi$  must satisfy,

$$X(t) = I + \int_{t_0}^t A(\tau)X(\tau)d\tau \quad (2.81)$$

$$\Phi(t, t_0) = I + \int_{t_0}^t A(\tau)\Phi(\tau, t_0)d\tau. \quad (2.82)$$

Subtracting and taking the norm, we get,

$$\|\Phi(t, t_0) - X(t)\| = \left\| \int_{t_0}^t A(\tau)(\Phi(\tau, t_0) - X(\tau))d\tau \right\| \quad (2.83)$$

$$\leq \left| \int_{t_0}^t \|A(\tau)\| \|\Phi(\tau, t_0) - X(\tau)\| d\tau \right|. \quad (2.84)$$

Applying the Gronwall lemma, we find that

$$\|\Phi(t, t_0) - X(t)\| = 0, \forall t \in \mathbb{R}. \quad (2.85)$$

We conclude that  $\Phi(t, t_0) = X(t)$  for all  $t \in \mathbb{R}$ , and that solutions to the IVP are unique.  $\square$

### 2.2.1.3 The State Transition Matrix

In the previous subsection, we developed the theory of the *Peano-Baker series* to prove the existence of a unique solution to the matrix initial value problem,

$$\dot{X}(t) = A(t)X(t), X(t_0) = I. \quad (2.86)$$

Since the integral formula for the Peano-Baker series is rather impractical to work with, we'll find it convenient to *abstract away* the computation of the Peano-Baker series and focus on  $\Phi(t, t_0)$  as the solution to the matrix initial value problem. In this spirit, we make the following definition.

**Definition 2.19 (State Transition Matrix)** Let  $A(\cdot) \in PC(\mathbb{R}, \mathbb{R}^{n \times n})$ . The state transition matrix with respect to  $A(\cdot)$  is a map  $\Phi : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}^{n \times n}$ , such that  $\Phi(\cdot, t_0) : \mathbb{R} \rightarrow \mathbb{R}^{n \times n}$  is the unique solution to the initial value problem,

$$\dot{X}(t) = A(t)X(t), X(t_0) = I. \quad (2.87)$$

*Remark 2.21* Despite its name, the state transition *matrix* is not a fixed matrix but rather a *map* into the set of matrices.



We emphasize—the state transition matrix  $\Phi(t, t_0)$  is *exactly* calculated by the Peano-Baker series. Here, we simply hide the Peano-Baker series behind a layer of abstraction—the *state transition matrix*—to emphasize that we don't want to use the series as an analysis tool.

By focusing on the “abstracted” definition of  $\Phi(t, t_0)$  as the unique solution of a differential equation, as opposed to the definition of  $\Phi(t, t_0)$  as an infinite series, we'll find that we can write much more elegant proofs.

**Proposition 2.6 (Properties of the State Transition Matrix)** *Let  $A(\cdot) \in \mathbb{PC}(\mathbb{R}, \mathbb{R}^{n \times n})$ . The state transition matrix  $\Phi$  with respect to  $A(\cdot)$  satisfies the following properties:*

1. *Composition:* For all  $t_0, t_1, t_2 \in \mathbb{R}$ ,  $\Phi(t_2, t_0) = \Phi(t_2, t_1)\Phi(t_1, t_0)$ .
2. *Inverse:* For all  $t_0, t_1 \in \mathbb{R}$ ,  $\Phi(t_1, t_0)$  is invertible with  $[\Phi(t_1, t_0)]^{-1} = \Phi(t_0, t_1)$ .

**Proof** First, we show the composition property. To prove this, we will use the uniqueness property of solutions to the initial value problem,  $\dot{X}(t) = A(t)X(t)$ ,  $X(t_0) = X_0$ , which follows from the Gronwall Lemma. In particular, we will show that, for all  $t_0, t_1 \in \mathbb{R}$ , both

$$\Phi(\cdot, t_0) : \mathbb{R} \rightarrow \mathbb{R}^{n \times n} \quad (2.88)$$

$$\Phi(\cdot, t_1)\Phi(t_1, t_0) : \mathbb{R} \rightarrow \mathbb{R}^{n \times n}, \quad (2.89)$$

are solutions to the matrix IVP  $\dot{X}(t) = A(t)X(t)$ ,  $X(t_1) = \Phi(t_1, t_0)$ . Then, we'll use uniqueness to conclude that they are equal. First, we know that  $\Phi(\cdot, t_0)$  is a solution to the matrix IVP by definition of the state transition matrix. Thus,  $\Phi(\cdot, t_0)$  satisfies,

$$\frac{d}{dt}\Phi(t, t_0) = A(t)\Phi(t, t_0) \quad \forall t \in \mathbb{R} \setminus D. \quad (2.90)$$

where  $D$  is the discontinuity set of  $A$ . This implies that  $\Phi(t, t_0)$  is also the solution the matrix IVP,  $\dot{X}(t) = A(t)X(t)$ ,  $X(t_1) = \Phi(t_1, t_0)$ . Now, we check the same for the second. We have that, for  $D$  the discontinuity set of  $A(\cdot)$ ,

$$\frac{d}{dt}[\Phi(t, t_1)\Phi(t_1, t_0)] = A(t)\Phi(t, t_1)\Phi(t_1, t_0) = A(t)[\Phi(t, t_1)\Phi(t_1, t_0)], \quad \forall t \in \mathbb{R} \setminus D. \quad (2.91)$$

Further, we have that  $\Phi(t_1, t_1)\Phi(t_1, t_0) = I\Phi(t_1, t_0) = \Phi(t_1, t_0)$ . Therefore,  $\Phi(\cdot, t_1)\Phi(t_1, t_0)$  *also* solves the initial value problem! By the Gronwall Lemma, it follows that solutions to the IVP are unique, which implies,

$$\Phi(t_2, t_1)\Phi(t_1, t_0) = \Phi(t_2, t_0), \quad \forall t_0, t_1, t_2 \in \mathbb{R}. \quad (2.92)$$

This completes the proof of the first item. The second item follows by direct application of the first. Fix times  $t_0, t_1 \in \mathbb{R}$ . Then, it follows from the composition rule that

$$\Phi(t_0, t_1)\Phi(t_1, t_0) = \Phi(t_0, t_0) = I \quad (2.93)$$

$$\Phi(t_1, t_0)\Phi(t_0, t_1) = \Phi(t_1, t_1) = I. \quad (2.94)$$

So, we conclude by the uniqueness of the matrix inverse that  $\Phi(t_0, t_1) = [\Phi(t_1, t_0)]^{-1}$ .  $\square$

### 2.2.1.4 The Continuous-Time, LTV Initial Value Problem

We're finally ready to tackle our original problem: proving the existence & uniqueness of solutions to the initial value problem,

$$\dot{x}(t) = A(t)x(t) + B(t)u(t), \quad x(t_0) = x_0. \quad (2.95)$$

Amazingly, *all we need* to construct solutions to this problem is the state transition matrix. To determine a formula for the solutions to such initial value problems, we'll need the differentiation under the integral rule, which we now recall.

**Theorem 2.4 (Leibniz Rule for Differentiation Under the Integral)** *Let  $f \in C^1(\mathbb{R} \times \mathbb{R}, \mathbb{R}^n)$  and  $a(\cdot), b(\cdot) \in C^1(\mathbb{R}, \mathbb{R})$  be continuously differentiable functions. For all  $t \in \mathbb{R}$ ,*

$$\frac{d}{dt} \left[ \int_{a(t)}^{b(t)} f(t, \tau) d\tau \right] = \int_{a(t)}^{b(t)} \frac{\partial}{\partial t} f(t, \tau) d\tau + f(t, b(t)) \frac{d}{dt} b(t) - f(t, a(t)) \frac{d}{dt} a(t). \quad (2.96)$$

**Proof** See Problem 2.6 for details.  $\square$

With this in mind, we state a theorem on the existence and uniqueness of solutions to the LTV initial value problem.

**Theorem 2.5 (Existence & Uniqueness of Solutions to LTV-IVP)** *Let  $A(\cdot) \in PC(\mathbb{R}, \mathbb{R}^{n \times n})$ ,  $B(\cdot) \in PC(\mathbb{R}, \mathbb{R}^{n \times m})$ , and  $u(\cdot) \in PC(\mathbb{R}, \mathbb{R}^m)$ . The unique solution to the initial value problem,*

$$\dot{x}(t) = A(t)x(t) + B(t)u(t), \quad x(t_0) = x_0, \quad (2.97)$$

*is given by the map  $x : \mathbb{R} \rightarrow \mathbb{R}^n$ , defined,*

$$x(t) = \Phi(t, t_0)x_0 + \int_{t_0}^t \Phi(t, \tau)B(\tau)u(\tau)d\tau, \quad (2.98)$$

*where  $\Phi$  is the state transition matrix with respect to  $A(\cdot)$ .*

**Proof** Our proof follows by direct differentiation. Let  $D$  be the shared discontinuity set of  $A(\cdot)$ ,  $B(\cdot)$ , and  $u(\cdot)$ . It follows from the uniform convergence property that  $\Phi(\cdot, \cdot)$  is continuously differentiable outside of the discontinuity set  $D$ . Let  $t \in \mathbb{R} \setminus D$ . By definition of the state transition matrix, it follows that

$$\frac{d}{dt}x(t) = \frac{d}{dt}\Phi(t, t_0)x_0 + \frac{d}{dt} \int_{t_0}^t \Phi(t, \tau)B(\tau)u(\tau)d\tau \quad (2.99)$$

$$= A(t)\Phi(t, t_0)x_0 + \Phi(t, t)B(t)u(t) + \int_{t_0}^t \frac{\partial}{\partial t} \Phi(t, \tau)B(\tau)u(\tau)d\tau \quad (2.100)$$

$$= A(t)\Phi(t, t_0)x_0 + B(t)u(t) + \int_{t_0}^t A(t)\Phi(t, \tau)B(\tau)u(\tau)d\tau \quad (2.101)$$

$$= A(t) \left[ \Phi(t, t_0)x_0 + \int_{t_0}^t \Phi(t, \tau)B(\tau)u(\tau)d\tau \right] + B(t)u(t) \quad (2.102)$$

$$= A(t)x(t) + B(t)u(t). \quad (2.103)$$

Note that here, since  $[t_0, t_1] \cap D$  contains a finite number of points (by definition of piecewise continuity), application of the Leibniz rule is justified. Thus, the solution satisfies the differentiation property. Also by definition of the state transition matrix,

$$x(t_0) = \Phi(t_0, t_0)x_0 + \int_{t_0}^{t_0} \Phi(t_0, \tau)B(\tau)u(\tau)d\tau = Ix_0 + 0 = x_0. \quad (2.104)$$

Therefore, the initial condition is also satisfied. This establishes the *existence* of solutions to the initial value problem. Now, we verify uniqueness using the Gronwall inequality. Suppose  $\hat{x} : \mathbb{R} \rightarrow \mathbb{R}^n$  is another solution to the initial value problem. Then, both  $x$  and  $\hat{x}$  must satisfy,

$$x(t) = x_0 + \int_{t_0}^t A(\tau)x(\tau) + B(\tau)u(\tau)d\tau \quad (2.105)$$

$$\hat{x}(t) = x_0 + \int_{t_0}^t A(\tau)\hat{x}(\tau) + B(\tau)u(\tau)d\tau, \quad \forall t \in \mathbb{R}. \quad (2.106)$$

Subtracting and taking the norm, one gets,

$$\|x(t) - \hat{x}(t)\| = \left\| \int_{t_0}^t A(\tau)(x(\tau) - \hat{x}(\tau))d\tau \right\| \quad (2.107)$$

$$\leq \left| \int_{t_0}^t \|A(\tau)\| \|x(\tau) - \hat{x}(\tau)\| d\tau \right|. \quad (2.108)$$

Applying the Gronwall inequality, it follows that  $\|x(t) - \hat{x}(t)\| = 0$ , for all  $t \in \mathbb{R}$ . We conclude that  $x = \hat{x}$ , and that solutions to the IVP are unique.  $\square$

This result yields the final piece in the puzzle in establishing that a continuous-time, linear time-varying system representation yields a continuous-time, linear I/O dynamical system. Since the existence and uniqueness theorem above does the bulk of the work, we leave the details of the following theorem to the reader.

**Theorem 2.6 (LTV System Representations Determine Linear I/O Systems)**

*Consider a continuous-time, LTV system representation  $(A(\cdot), B(\cdot), C(\cdot), D(\cdot))$ . This representation determines a linear I/O system  $\mathcal{D}$ , comprised of the following data:*

1. Time Set:  $\mathcal{T} = \mathbb{R}$ .
2. Spaces:  $\mathcal{U} = PC(\mathbb{R}, \mathbb{R}^m)$ ,  $\mathcal{Y} = PC(\mathbb{R}, \mathbb{R}^p)$ , and  $\Sigma = \mathbb{R}^n$ .
3. State Transition Map: the state transition map is computed,

$$\varphi(t_1, t_0, x_0, u(\cdot)) = \Phi(t_1, t_0)x_0 + \int_{t_0}^{t_1} \Phi(t_1, \tau)B(\tau)u(\tau)d\tau. \quad (2.109)$$

4. Readout Map: the readout map is computed,

$$r(t, x, u) = C(t)x + D(t)u. \quad (2.110)$$

5. I/O Map: the I/O map is computed,

$$\rho(t_1, t_0, x_0, u(\cdot)) = C(t_1)\Phi(t_1, t_0)x_0 + C(t_1) \int_{t_0}^{t_1} \Phi(t_1, \tau)B(\tau)u(\tau)d\tau + D(t_1)u(t_1). \quad (2.111)$$

**Proof** See Problem 2.7. □

To conclude, we make the following important observation. The evolution of the output of any continuous-time, LTV system representation can be *decomposed* into the sum of the zero-input and zero-state components:

$$\rho(t_1, t_0, x_0, u(\cdot)) = \underbrace{C(t_1)\Phi(t_1, t_0)x_0}_{\text{Zero-Input Response}} + \underbrace{C(t_1) \int_{t_0}^{t_1} \Phi(t_1, \tau)B(\tau)u(\tau)d\tau + D(t_1)u(t_1)}_{\text{Zero-State Response}}. \quad (2.112)$$

Note that the zero-input and zero-state responses are sometimes referred to as the *free* and *forced* responses, respectively. This tells us that the response of any linear, time-varying system to an input signal has a component due to the initial condition and a separate component due to the input.

## 2.2.2 Solutions of Discrete-Time Linear Systems

After taking on a monumental challenge in the desert of continuous-time systems, it's now time to relax in the oasis of discrete-time systems. Kick back, grab your favorite normed vector space, and prepare to be relieved by a *substantially* easier theory.

In this section, we'll work through the process of proving that discrete-time, linear time-varying systems define discrete-time linear I/O systems. Why is the theory in the discrete-time case so much easier than in the continuous-time case? Let's take a quick look at the state equation for the discrete-time, linear time-varying system and see what we find. We have,

$$x[k+1] = A[k]x[k] + B[k]u[k]. \quad (2.113)$$

That is, given  $x[k]$  and  $u[k]$ , we can immediately calculate  $x[k+1] = A[k]x[k] + B[k]u[k]$ . This means that, for any initial condition  $x[k_0] = x_0$  and input signal  $u[\cdot] : \mathbb{Z} \rightarrow \mathbb{R}^m$ , we can recursively solve for  $x[k]$ ,  $k \geq k_0$ . This makes the problem of existence and uniqueness of solutions trivial in discrete-time.

Despite this great simplification of the discrete-time initial value problem, we'll still find it fruitful to examine closely the structure of solutions to a discrete-time system—just because we can write down a solution directly from the recurrence doesn't mean there isn't more at play! Interestingly, we'll find that a state transition matrix similar to that of the continuous-time case also appears in the discrete-time case.

In the remainder of this section, we'll follow the same general procedure as in the continuous-time case. Here, we'll leave many of the results as exercises or problems, due to their simpler analytical nature.

### 2.2.2.1 Defining Solutions to Discrete-Time Systems

As with the continuous-time case, we begin by specifying a formal definition of a solution to a discrete-time initial value problem. In this case, since the time set is discrete and the state equation is a recurrence relation, we won't need to worry about regularity conditions such as piecewise continuity. This is reflected in the simpler form of the definition.

**Definition 2.20 (Solution to Discrete-Time Recurrence)** Let  $A[\cdot] : \mathbb{Z} \rightarrow \mathbb{R}^{n \times n}$ ,  $B[\cdot] : \mathbb{Z} \rightarrow \mathbb{R}^{n \times m}$ , and  $u[\cdot] : \mathbb{Z} \rightarrow \mathbb{R}^m$  be sequences. For  $x_0 \in \mathbb{R}^n$  and  $k_0 \in \mathbb{Z}$ , a solution to the discrete-time recurrence,

$$x[k+1] = A[k]x[k] + B[k]u[k], \quad x[k_0] = x_0, \quad (2.114)$$

is a sequence  $x[\cdot] : \mathbb{Z}_{\geq k_0} \rightarrow \mathbb{R}^n$ , satisfying:

1. Initial condition:  $x[k_0] = x_0$ .
2. Recurrence: For all  $k \geq k_0$ ,  $x[k+1] = A[k]x[k] + B[k]u[k]$ .

We observe that the solution to a discrete-time initial value problem has exactly the structure we expect! It's important to note that—instead of the solution being defined on all of  $\mathbb{Z}$ , solutions are defined as sequences starting at a time  $k_0$ .

The reasoning behind this is essentially as follows: for  $k < k_0$ , it's possible that the trajectory leading to  $x[k_0] = x_0$  is *not* uniquely defined (can you think of the reason why? We'll provide an answer below). For these reasons, we restrict the definition of a solution to be for  $k \geq k_0$ . Shortly, we'll describe some conditions that let us extend the domain of the definition of a solution to all of  $\mathbb{Z}$ .

### 2.2.2.2 A Matrix Initial Value Problem & The State Transition Matrix

In order to uncover the structure underlying the discrete-time initial value problem, we again take the approach of studying a *matrix* initial value problem. Fortunately, as mentioned above, the existence and uniqueness of solutions are no longer a concern! As such, we can directly jump to the definition of a state transition matrix.

**Definition 2.21 (Discrete-Time State Transition Matrix)** Consider a sequence  $A[\cdot] : \mathbb{Z} \rightarrow \mathbb{R}^{n \times n}$ . The (discrete-time) state transition matrix with respect to  $A[\cdot]$  is a map

$$\Phi[\cdot, \cdot] : \mathbf{T} \rightarrow \mathbb{R}^{n \times n}, \quad \mathbf{T} := \{(k, k_0) \in \mathbb{Z} \times \mathbb{Z} : k \geq k_0\}, \quad (2.115)$$

such that for all  $k_0 \in \mathbb{Z}$ ,  $\Phi[\cdot, k_0]$  is the solution to the discrete-time, matrix initial value problem

$$X[k+1] = A[k]X[k], \quad X[k_0] = I. \quad (2.116)$$

Thus, we define the state transition matrix  $\Phi$  in *exactly* the same way as for the continuous-time case—as a solution to a matrix initial value problem defined by  $A[\cdot]$ , with an initial condition given by the identity matrix. Here, however, in order to get a well-defined solution, we must define the domain of  $\Phi$  such that  $k \geq k_0$ . We'll see shortly how this assumption on the domain can be relaxed when the sequence  $\{A[k]\}_{k \in \mathbb{Z}}$  has all nonsingular elements.

**Proposition 2.7 (Structure of the Discrete-Time State Transition Matrix)** *Consider a sequence  $A[\cdot] : \mathbb{Z} \rightarrow \mathbb{R}^{n \times n}$ . For any  $k_0 \in \mathbb{Z}$  and  $k \geq k_0$ , the state transition matrix  $\Phi[k, k_0]$  with respect to  $A[\cdot]$  is computed*

$$\Phi[k_0, k_0] = I \quad (2.117)$$

$$\Phi[k+1, k_0] = A[k]\Phi[k, k_0]. \quad (2.118)$$

**Exercise 2.6** Prove Proposition 2.7.

This result gives a method of calculating the state transition matrix for a given  $k_0$  and all  $k \geq k_0$ . Why can't we calculate the state transition matrix for all  $k \in \mathbb{Z}$ ? In the event where the matrix  $A[k]$  is not invertible, we lose uniqueness in the definition of  $\Phi$ —thus, for  $k < k_0$ ,  $\Phi$  might be ill-defined. In the event where  $A[k]$  is nonsingular (i.e. invertible), however, we can make the following conclusion.

**Proposition 2.8 (State Transition Matrix for Invertible  $A$ )** *Consider a sequence  $A[\cdot] : \mathbb{Z} \rightarrow \mathbb{R}^{n \times n}$ , in which each  $A[k]$  is nonsingular. For such a sequence, the state transition matrix  $\Phi$  can be uniquely defined on all of  $\mathbb{Z} \times \mathbb{Z}$ .*

**Exercise 2.7** Prove Proposition 2.8. What can go wrong if  $A[k]$  is singular? Provide an example.

Finally, we show that the discrete-time state transition matrix satisfies a composability property. Here, due to the risk of singular  $A[k]$ , we *cannot* prove a composability property for all  $k_0, k_1, k_2$ —we are restricted to  $k_0 \leq k_1 \leq k_2$ . For this same reason, we are not guaranteed that the discrete-time state transition matrix is invertible for all  $k_0, k_1 \in \mathbb{Z}$ .

**Proposition 2.9 (Composability of the Discrete-Time State Transition Matrix)** *Consider a sequence  $A[\cdot] : \mathbb{Z} \rightarrow \mathbb{R}^{n \times n}$ . The state transition matrix  $\Phi$  with respect to  $A[\cdot]$  then satisfies  $\Phi[k_2, k_0] = \Phi[k_2, k_1]\Phi[k_1, k_0]$ , for all  $k_0 \leq k_1 \leq k_2 \in \mathbb{Z}$ .*

### 2.2.2.3 Solutions to the Discrete-Time, LTV Recurrence

Using the state transition matrix, we can find the unique solution to the discrete-time recurrence defined by the state equation of the discrete-time, LTV representation.

**Theorem 2.7 (Solutions to Discrete-Time, LTV Recurrence)** *Let  $A[\cdot] : \mathbb{Z} \rightarrow \mathbb{R}^{n \times n}$ ,  $B[\cdot] : \mathbb{Z} \rightarrow \mathbb{R}^{n \times m}$ , and  $u[\cdot] : \mathbb{Z} \rightarrow \mathbb{R}^m$  be sequences. For  $x_0 \in \mathbb{R}^n$  and  $k_0 \in \mathbb{Z}$ , the unique solution to the discrete-time recurrence,*

$$x[k+1] = A[k]x[k] + B[k]u[k], \quad x[k_0] = x_0, \quad (2.119)$$

*is given by the sequence  $x[\cdot] : \mathbb{Z}_{\geq k_0} \rightarrow \mathbb{R}^n$ , defined,*

$$x[k] = \Phi[k, k_0]x_0 + \sum_{j=k_0}^{k-1} \Phi[k, j+1]B[j]u[j]. \quad (2.120)$$

**Exercise 2.8** Prove Theorem 2.7 by induction on  $k$ . Why is the expression  $\Phi[k, j+1]$  well-defined for  $j \in [k_0, k-1] \cap \mathbb{Z}$ ?

Finally, we confirm that discrete-time, linear time-varying representations determine discrete-time linear I/O systems.

**Theorem 2.8 (DT-LTV System Representations Determine DT Linear I/O Systems)** *Consider a discrete-time, LTV system representation  $(A[\cdot], B[\cdot], C[\cdot], D[\cdot])$ . This representation determines a discrete-time linear I/O system  $\mathcal{D}$ , comprised of the data:*

1. Time Set:  $\mathcal{T} = \mathbb{Z}$ .
2. Spaces:  $\mathcal{U} = \{u : \mathbb{Z} \rightarrow \mathbb{R}^m\}$ ,  $\mathcal{Y} = \{y : \mathbb{Z} \rightarrow \mathbb{R}^p\}$ , and  $\Sigma = \mathbb{R}^n$ .
3. State Transition Map: the state-transition map is computed,

$$\varphi(k_1, k_0, x_0, u[\cdot]) = \Phi[k_1, k_0]x_0 + \sum_{j=k_0}^{k_1-1} \Phi[k_1, j+1]B[j]u[j] \quad (2.121)$$

4. Readout Map: the readout map is computed,

$$r(k, x, u) = C[k]x + D[k]u \quad (2.122)$$

5. I/O Map: The I/O map is computed,

$$\rho(k_1, k_0, x_0, u[\cdot]) = C[k_1]\Phi[k_1, k_0]x_0 + C[k_1] \sum_{j=k_0}^{k_1-1} \Phi[k_1, j+1]B[j]u[j] + D[k_1]u[k_1]. \quad (2.123)$$

**Proof** See Problem 2.7. □

As with the case of a continuous-time linear system, we note that the state response of any discrete-time linear, time-varying system representation can be decomposed as the sum,

$$\rho(k_1, k_0, x_0, u[\cdot]) = \underbrace{C[k_1]\Phi[k_1, k_0]x_0}_{\text{Zero-Input Response}} + \underbrace{C[k_1] \sum_{j=k_0}^{k_1-1} \Phi[k_1, j+1]B[j]u[j] + D[k_1]u[k_1]}_{\text{Zero-State Response}}, \quad (2.124)$$

of a zero-input and a zero-state response. As with the continuous-time case, these components are also referred to as the *free* and *forced* responses, respectively.

### 2.2.3 Further Reading

This section was mainly influenced by [8], [25], and [27]. For an approach to the existence & uniqueness problem that uses a more general existence & uniqueness theorem for differential equations, the interested reader is encouraged to consult [20]. For a treatment of existence & uniqueness of solutions to differential equations with *measurable* data (more general than piecewise continuous), a measure-theoretic treatment of ordinary differential equations is found in Appendix C of [36].

### 2.2.4 Problems

**Problem 2.3 (Transition Matrix Under Change of Variable)** Consider a continuous-time linear, time-varying system representation  $(A(\cdot), B(\cdot), C(\cdot), D(\cdot))$ ,

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) \quad (2.125)$$

$$y(t) = C(t)x(t) + D(t)u(t). \quad (2.126)$$

1. Consider an invertible linear transformation  $T \in \mathbb{R}^{n \times n}$  and a corresponding change of variables,  $z = Tx$ . Identify the system representation  $(\hat{A}(\cdot), \hat{B}(\cdot), \hat{C}(\cdot), \hat{D}(\cdot))$  for which solutions to,

$$\dot{z}(t) = \hat{A}(t)\hat{z}(t) + \hat{B}(t)u(t) \quad (2.127)$$

$$\hat{y}(t) = \hat{C}(t)\hat{z}(t) + \hat{D}(t)u(t) \quad (2.128)$$

satisfy  $z(t) = Tx(t)$  and  $\hat{y}(t) = y(t)$  for all initial conditions  $x_0$  and  $Tx_0$  and piecewise continuous input signals  $u(\cdot)$ . Conclude that the input to output behavior of the system *does not* depend on changes of state coordinates.

2. Write the state transition matrix  $\hat{\Phi}(t, t_0)$  of the transformed system in terms of the state transition matrix  $\Phi(t, t_0)$  of the original system and the transformation  $T$ .
3. Does the relation you derived in part (2) also hold for a discrete-time system representation? Explain why or why not.

**Problem 2.4 (The Intermediate Value Property of the Derivative)** In this problem, we formalize the “intermediate value property” of the derivative, which states that the derivative of a function cannot have any jump discontinuities. Recall that the derivative of a scalar function  $f : \mathbb{R} \rightarrow \mathbb{R}$  is defined,

$$f'(t) = \lim_{\tau \rightarrow t} \frac{f(t) - f(\tau)}{t - \tau}. \quad (2.129)$$

It is *not* necessarily the case that the derivative of a differentiable function is continuous! However, we *can* exclude certain types of discontinuities.

1. First, suppose  $f : \mathbb{R} \rightarrow \mathbb{R}$  is differentiable on an open interval  $(a, b)$ . Show that if  $f$  attains a maximum or minimum value at a point  $c \in (a, b)$ , then  $f'(c) = 0$ .
2. Now, suppose  $f : \mathbb{R} \rightarrow \mathbb{R}$  is differentiable on an open set  $A \subseteq \mathbb{R}$  containing an interval  $[a, b]$ . Suppose  $\alpha \in \mathbb{R}$  satisfies  $f'(a) < \alpha < f'(b)$ . Show there exists a point  $c \in (a, b)$  for which  $f'(c) = \alpha$ . *If you get stuck, consult Chapter 5.2 of [1] for some hints.*
3. Apply the conclusion of part (2) to study the jump discontinuities of the derivative of a function  $f : \mathbb{R} \rightarrow \mathbb{R}^n$ .

**Problem 2.5 (Gronwall Inequality)** In this problem, we’ll walk through a proof of the Gronwall inequality (Lemma 2.3). Recall that the Gronwall inequality is formulated as follows. Let  $y, k \in PC(\mathbb{R}, \mathbb{R}_{\geq 0})$  and  $\mu \in PC(\mathbb{R}, \mathbb{R}_{\geq 0})$ ,  $c \in \mathbb{R}_{\geq 0}$ , and  $t_0 \in \mathbb{R}$ . If for all  $t \in \mathbb{R}$ ,  $y$  satisfies,

$$y(t) \leq c + \left| \int_{t_0}^t k(\tau)y(\tau)d\tau \right|, \quad (2.130)$$



then for all  $t \in \mathbb{R}$ ,

$$y(t) \leq c \exp \left| \int_{t_0}^t k(\tau) d\tau \right|. \quad (2.131)$$

Let's get to work on assembling a proof of this result.

1. Fix times  $t, t_0 \in \mathbb{R}$  with  $t > t_0$ . Define a function,

$$Y(t) = c + \int_{t_0}^t k(\tau) y(\tau) d\tau. \quad (2.132)$$

Argue that  $y(t) \leq Y(t)$  for all  $t \geq t_0$ , and that  $Y(t)$  satisfies  $\frac{d}{dt}Y(t) = k(t)y(t)$ .

2. Prove that,

$$y(t) \leq Y(t)k(t) \exp\left(-\int_{t_0}^t k(\tau) d\tau\right), \quad (2.133)$$

and that

$$\frac{d}{dt}[Y(t) \exp(-\int_{t_0}^t k(\tau) d\tau)] \leq 0. \quad (2.134)$$

3. Conclude that  $y(t) \leq Y(t) \leq ce^{\int_{t_0}^t k(\tau) d\tau}$ .

**Problem 2.6 (Differentiation Under the Integral Sign)** Using the limit definition of the derivative, prove Theorem 2.4, the Leibniz rule for differentiation under the integral sign.

**Problem 2.7 (LTV System Representations Determine Linear I/O Systems)**

Above, we stated two Theorems - 2.6 and 2.8 - which claimed that linear time-varying system representations generate linear I/O dynamical systems. Supply proofs of Theorems 2.6 and 2.8.

**Problem 2.8 (An Inverse Initial Value Problem)** We know that  $\Phi(t, t_0)$  is the solution to the initial value problem  $\dot{X}(t) = A(t)X(t)$ ,  $X(t_0) = I$ . In this problem, we'll find out what  $\Phi(t_0, t)$  corresponds to.

1. Consider a continuously differentiable, matrix-valued function  $M(\cdot) : \mathbb{R} \rightarrow \mathbb{R}^{n \times n}$ . Suppose for all  $t \in \mathbb{R}$ ,  $M(t)$  is nonsingular. Determine an expression for  $\frac{d}{dt}[M^{-1}(t)]$  in terms of  $\dot{M}(t)$  and  $M^{-1}(t)$ .
2. Now, consider a matrix  $A(\cdot) \in PC(\mathbb{R}, \mathbb{R}^{n \times n})$ . Find an expression for the derivative  $\frac{\partial}{\partial \tau}\Phi(t, \tau)$  of the state transition matrix  $\Phi$  with respect to  $A(\cdot)$ , in terms of  $\Phi(t, t_0)$  and  $A(t)$ . *You may assume the derivative is being taken at a point where  $A(\cdot)$  is continuous.*
3. Prove that  $\Phi(t_0, t)$  is the unique solution of the matrix initial value problem,

$$\dot{X}(t) = -X(t)A(t), \quad X(t_0) = I. \quad (2.135)$$

**Problem 2.9 (The Jacobi-Liouville Formula ★ [10])** Above, we showed that the continuous-time state transition matrix is always invertible. Here, we'll provide another proof of this by means of the *Jacobi-Liouville formula*, which explicitly provides a formula

for the determinant of the state transition matrix. In particular, the Jacobi-Liouville formula is,

$$\det \Phi(t, t_0) = \exp \left( \int_{t_0}^t \operatorname{tr}(A(\tau)) d\tau \right). \quad (2.136)$$

1. Prove that, for  $M \in \mathbb{R}^{n \times n}$  and  $\epsilon \in \mathbb{R}$ , there exists a continuous function  $R : \mathbb{R} \rightarrow \mathbb{R}$  for which

$$\det(I + \epsilon M) = 1 + \epsilon \operatorname{tr}(M) + R(\epsilon) \text{ and } \lim_{\epsilon \rightarrow 0} \frac{R(\epsilon)}{\epsilon} = 0. \quad (2.137)$$

*Hint: consider working with eigenvalues.*

2. Using the determinant formula from (1), show that

$$\frac{d}{dt} \det[\Phi(t, t_0)] = \operatorname{tr}(A(t)) \det[\Phi(t, t_0)]. \quad (2.138)$$

*Hint: Work with the limit definition of the derivative. If you use a Taylor approximation, be rigorous about your use of the remainder term.*

3. Conclude the Jacobi-Liouville formula. Using the Jacobi-Liouville formula, provide a proof that  $\Phi(t, t_0)$  is invertible for all  $(t, t_0) \in \mathbb{R} \times \mathbb{R}$ .

**Problem 2.10 (Solution of a Matrix Differential Equation [9])** Let  $A_1(\cdot)$ ,  $A_2(\cdot)$ , and  $F(\cdot)$  be elements of  $PC(\mathbb{R}, \mathbb{R}^{n \times n})$ . Let  $\Phi_i$  be the state transition matrix of  $\dot{x}(t) = A_i(t)x(t)$  for  $i = 1, 2$ . Show that the solution of the matrix differential equation:

$$\dot{X}(t) = A_1(t)X(t) + X(t)A_2^\top(t) + F(t), \quad X(t_0) = X_0, \quad (2.139)$$

is given by,

$$X(t) = \Phi_1(t, t_0)X_0\Phi_2^\top(t, t_0) + \int_{t_0}^t \Phi_1(t, \tau)F(\tau)\Phi_2^\top(t, \tau)d\tau. \quad (2.140)$$

Is this the unique solution of the matrix differential equation? Back up your answer with a proof or disproof.

**Problem 2.11 (A Special State Transition Matrix)** Consider a piecewise continuous matrix  $A \in PC(\mathbb{R}, \mathbb{R}^{n \times n})$ , and let  $\Phi$  denote the state transition matrix of  $\dot{x}(t) = A(t)x(t)$ . If for every  $(\tau, t) \in \mathbb{R} \times \mathbb{R}$ , one has,

$$A(t) \left( \int_{\tau}^t A(\eta) d\eta \right) = \left( \int_{\tau}^t A(\eta) d\eta \right) A(t), \quad (2.141)$$

prove using the Peano-Baker series that,

$$\Phi(t, \tau) = \exp \left( \int_{\tau}^t A(\eta) d\eta \right) = \sum_{k=0}^{\infty} \frac{1}{k!} \left( \int_{\tau}^t A(\eta) d\eta \right)^k. \quad (2.142)$$

Using this result, calculate the state transition matrix associated to the matrix,

$$A(t) = \begin{bmatrix} 0 & 0 \\ t & 0 \end{bmatrix}. \quad (2.143)$$

## 2.3 Solutions of Linear, Time-Invariant Systems

In the previous section, we studied the structure of solutions to linear, time-varying systems. Now, we'll specialize to the linear, time-*invariant* case. Since the set of linear, time-invariant systems are a *strict subset* of the set of linear, time-varying systems, we can directly apply all of the results we developed in the previous section to the LTI case. Thus, we immediately have access to all of our results concerning the state transition matrix, existence & uniqueness of solutions, the structure of solutions, and so on.

What, then, do we aim to accomplish by focusing on the linear, time-invariant case? Recall that in the previous section, we established that the state transition matrix of a linear, time-varying continuous-time system can be computed via the *Peano-Baker series*,

$$\Phi(t, t_0) = I + \int_{t_0}^t A(\tau) d\tau + \int_{t_0}^t A(\tau) \int_{t_0}^{\tau} A(\tau') d\tau' d\tau + \dots, \quad (2.144)$$

a rather unwieldy infinite series with no immediately apparent simplifications. Likewise, we showed that the state transition matrix of a linear, time-varying discrete-time system is computed via the similarly unrevealing product,

$$\Phi[k, k_0] = A[k-1]A[k-2] \cdot \dots \cdot A[k_0+1]A[k_0]. \quad (2.145)$$

Let's study how these expressions simplify in the time-invariant case. We recall that in the time-invariant case, the matrices  $A(\cdot)$  and  $A[\cdot]$  cease to be functions of time, and are specified by a constant matrix  $A \in \mathbb{R}^{n \times n}$ . Let's substitute such a matrix into the Peano-Baker series. A little bit of computation yields,

$$\Phi(t, t_0) = I + A(t - t_0) + \frac{A^2(t - t_0)^2}{2} + \frac{A^3(t - t_0)^3}{6} + \dots \quad (2.146)$$

Interestingly, such an expression seems to mirror the power series,

$$\sum_{k=1}^{\infty} \frac{A^k(t - t_0)^k}{k!}. \quad (2.147)$$

If  $A$  were a scalar,  $A = a \in \mathbb{R}$ , this would mean that the Peano-Baker series would *exactly* equal the exponential,  $\exp(a(t - t_0))$ —amazing! Now, we examine the discrete-time case. Here, the product  $A[k-1]A[k-2] \cdot \dots \cdot A[k_0]$  simply reduces to the matrix power  $A^{k-k_0}$ .

This simple analysis leads us to the following conclusion: compared to the time-varying case, the time-invariant case has a *significant* amount of structure that we can exploit. In particular, by examining properties of power series of matrices (in the continuous-time case) and of exponents of matrices (in the discrete-time case), we can gain significant insight into the behavior of linear, time-invariant systems. Let's begin!

### 2.3.1 Continuous-Time LTI Systems

We begin our study of linear, time-invariant systems with the continuous-time case. In our brief, expository analysis above, we discovered that the Peano-Baker series formula for the

state transition matrix *seems* to follow the pattern,

$$\Phi(t, t_0) = I + A(t - t_0) + \frac{A^2(t - t_0)^2}{2} + \frac{A^3(t - t_0)^3}{6} + \dots \quad (2.148)$$

That is, it appeared as if the state transition matrix for a continuous-time, linear time-invariant system was computable from the power series for the exponential. Is this truly the case? Let's find out! As a first step, we formally define the exponential of a matrix. For the sake of generality, we make this definition for a *complex* matrix.

**Definition 2.22 (Matrix Exponential)** The matrix exponential on  $\mathbb{C}^{n \times n}$  is the mapping  $\exp : \mathbb{C}^{n \times n} \rightarrow \mathbb{C}^{n \times n}$ , mapping a matrix  $A \in \mathbb{C}^{n \times n}$  to  $\exp(A)$ ,

$$\exp(A) := I + A + \frac{A^2}{2!} + \frac{A^3}{3!} + \dots = \sum_{k=0}^{\infty} \frac{A^k}{k!}. \quad (2.149)$$

*Remark 2.22* Here, we'll denote the matrix exponential of  $A \in \mathbb{C}^{n \times n}$  either by  $\exp(A)$  or  $e^A$ . The choice between the two is really a matter of personal preference.

In order to ensure the matrix exponential is *well-defined*, we must ensure that it converges. For the case of  $A \in \mathbb{R}^{n \times n}$ , this is something we've already proven! Recall that in the previous section, we proved that the Peano-Baker series converges uniformly on any interval  $[-t', t']$ , where  $t' \in \mathbb{R}_{\geq 0}$ .

Since the exponential of a real matrix is a special case of the Peano-Baker series, for which  $A(\cdot) = A \in \mathbb{R}^{n \times n}$ , the exponential of a real matrix enjoys similar convergence properties to the Peano-Baker series. Luckily, none of the convergence properties are lost when generalizing from  $A \in \mathbb{R}^{n \times n}$  to  $A \in \mathbb{C}^{n \times n}$ . We state this fact in the following proposition.

**Proposition 2.10 (Convergence of the Matrix Exponential)** *Let  $A \in \mathbb{C}^{n \times n}$ . On any compact interval  $[-t', t'] \subseteq \mathbb{R}$ ,  $t' \in \mathbb{R}_{\geq 0}$ , the power series,*

$$\exp(At) = \sum_{k=0}^{\infty} \frac{A^k t^k}{k!}, \quad (2.150)$$

*converges uniformly.*

*Remark 2.23* By taking  $t' \geq 1$  and  $t = 1$ , one may confirm by application of Proposition 2.10 that the exponential  $\exp(A)$  converges for any  $A \in \mathbb{C}^{n \times n}$ .

**Exercise 2.9** Prove Proposition 2.10 directly without use of the Peano-Baker series. Note that the Weierstrass M-test *still holds* on complex-valued, finite-dimensional vector spaces!

Now that we've established its convergence, we know that the matrix exponential is a well-defined quantity. Thus, we can move on to study other properties of the exponential. In the following proposition, we summarize a few important, algebraic properties of the matrix exponential.

**Proposition 2.11 (Basic Properties of the Matrix Exponential)** *Consider a matrix  $A \in \mathbb{C}^{n \times n}$ . The exponential of  $A$  satisfies the following algebraic properties:*

1. Eigenvalue-eigenvector pairs: for an eigenvalue-eigenvector pair  $(\lambda, v)$  of  $A$ ,  $(e^\lambda, v)$  is an eigenvalue-eigenvector pair of  $\exp(A)$ .

2. *Determinant*: The determinant of the exponential is computed  $\det(\exp A) = e^{\text{tr } A}$ .  
 3. *Invertibility*: The matrix exponential is invertible, with  $(\exp(A))^{-1} = \exp(-A)$ .

**Proof** See Problem 2.14. □

Now that we've defined the matrix exponential and some of its basic, algebraic properties, we consider its applications in the study of linear, time-invariant systems. Recall that—when motivating the matrix exponential—we derived the formula for the exponential by substituting a constant matrix into the formula for the Peano-Baker series. Since the Peano-Baker series is used to compute the state transition matrix of a linear, time-varying system, it stands to reason that the matrix exponential can be used to compute the state transition matrix of a linear, time-invariant system.

**Proposition 2.12 (Continuous-Time, LTI State Transition Matrix)** *Consider a fixed matrix  $A \in \mathbb{R}^{n \times n}$ . The continuous-time state transition matrix with respect to  $A$  is computed,*

$$\Phi(t, t_0) = \exp(A(t - t_0)). \quad (2.151)$$

*That is,  $\Phi(t, t_0) = \exp(A(t - t_0))$  is the unique solution to the matrix initial value problem,*

$$\dot{X}(t) = AX(t), \quad X(t_0) = I. \quad (2.152)$$

*Remark 2.24* In this initial value problem,  $A(\cdot) = A$  is constant. Since constant functions are *always continuous*, the discontinuity set of  $A(\cdot) = A$  is empty. Thus, we must verify that the derivative condition  $\dot{X}(t) = AX(t)$  holds for all  $t \in \mathbb{R}$ .

**Proof** There are a few ways we can prove this result. First, we could verify that the exponential satisfies the initial value problem by showing the exponential  $\exp(A(t - t_0))$  equals the Peano-Baker series. Secondly, we can proceed by direct differentiation. Here, we'll take the direct differentiation approach in order to get a feel for the definition of the exponential—we leave the (more straightforward) Peano-Baker series approach as an exercise below.

Let's show via direct differentiation that  $\exp(A(t - t_0))$  is the solution to the given initial value problem. First, let's examine the derivatives of the partial sums of the exponential. Fix a pair of times  $t, t_0 \in \mathbb{R}$ . We have, for  $p \in \mathbb{Z}_{\geq 0}$ ,

$$\frac{d}{dt} \sum_{k=0}^p \frac{A^k(t - t_0)^k}{k!} = \sum_{k=0}^p \frac{d}{dt} \frac{A^k(t - t_0)^k}{k!} = \sum_{k=1}^p \frac{A^k(t - t_0)^{k-1}}{(k-1)!} = A \sum_{k=0}^{p-1} \frac{A^k(t - t_0)^k}{k!}. \quad (2.153)$$

Let's use uniform convergence to pass to the limit. Fix a time  $t' \geq 0$  for which  $t, t_0 \in (-t', t')$ . We know that  $\exp(A(t - t_0))$  converges uniformly on  $[-t', t']$ , and that the sequence of derivatives above converges uniformly on  $[-t', t']$ . This implies,

$$\frac{d}{dt} \lim_{p \rightarrow \infty} \sum_{k=0}^p \frac{A^k(t - t_0)^k}{k!} = \lim_{p \rightarrow \infty} \sum_{k=0}^p \frac{d}{dt} \frac{A^k(t - t_0)^k}{k!} = A \sum_{k=0}^{\infty} \frac{A^k(t - t_0)^k}{k!}. \quad (2.154)$$

We conclude that  $\frac{d}{dt} \exp(A(t - t_0)) = A \exp(A(t - t_0))$ . Additionally, by definition of the power series, we have  $\exp(A(t_0 - t_0)) = \exp(0) = I$ . Thus,  $\exp(A(t - t_0))$  is a solution to the initial value problem,

$$\dot{X}(t) = AX(t), \quad X(t_0) = I. \quad (2.155)$$

By uniqueness of solutions, it follows that  $\Phi(t, t_0) = \exp(A(t - t_0))$ .  $\square$

**Exercise 2.10** Provide an alternate proof of Proposition 2.12 by showing the matrix exponential  $\exp(A(t - t_0))$  equals the Peano-Baker series defined by  $A$ .

Now that we've confirmed that the matrix exponential enables computation of the state transition matrix in the linear, time-invariant case, all of the properties we proved about state transition matrices in the previous section immediately apply to the exponential. With this knowledge, we confirm that a continuous-time, LTI system representation determines a *formal* continuous-time, LTI dynamical system.

**Theorem 2.9 (CT-LTI Representation Determines a CT-LTI System)** *Consider a continuous-time, LTI system representation  $(A, B, C, D)$ . This representation determines a continuous-time, linear time-invariant I/O system  $\mathcal{D}$ , specified by the following data:*

1. *Time set:*  $\mathcal{T} = \mathbb{R}$ .
2. *Spaces:*  $\mathcal{U} = PC(\mathbb{R}, \mathbb{R}^m)$ ,  $\mathcal{Y} = PC(\mathbb{R}, \mathbb{R}^p)$ , and  $\Sigma = \mathbb{R}^n$ .
3. *State transition map:* the state transition map is computed,

$$\varphi(t_1, t_0, x_0, u(\cdot)) = e^{A(t_1 - t_0)}x_0 + \int_{t_0}^{t_1} e^{A(t_1 - \tau)}Bu(\tau)d\tau. \quad (2.156)$$

4. *Readout map:* the readout map is computed,

$$r(t, x, u) = Cx + Du \quad (2.157)$$

5. *I/O map:* The I/O map is computed,

$$\rho(t_1, t_0, x_0, u(\cdot)) = Ce^{A(t_1 - t_0)}x_0 + C \int_{t_0}^{t_1} e^{A(t_1 - \tau)}Bu(\tau)d\tau + Du(t_1). \quad (2.158)$$

**Proof** The proof of this result follows directly from application of Theorem 2.6, Proposition 2.12, and verification of the time-invariance property.  $\square$

**Exercise 2.11** Provide the details of the proof of Theorem 2.9.

### 2.3.2 Discrete-Time LTI Systems

Now, we undertake a similar procedure for the discrete-time case. As with the continuous-time case, we begin by computing the state transition matrix for a discrete-time, linear time-invariant system.

**Proposition 2.13 (Discrete-Time, LTI State Transition Matrix)** *Consider a fixed matrix  $A \in \mathbb{R}^{n \times n}$ . The discrete-time state transition matrix with respect to  $A$  is computed,*

$$\Phi[k, k_0] = A^{k - k_0}, \quad \forall k \geq k_0. \quad (2.159)$$

*This is the unique solution to the matrix recurrence,  $X[k + 1] = AX[k]$ ,  $X[k_0] = I$ .*

**Exercise 2.12** Prove Proposition 2.13 by using the formula for the discrete-time, linear time-varying state transition matrix derived in the previous section.

Using this result, we can prove that any discrete-time, LTI system representation induces a *formal* discrete-time, LTI I/O dynamical system. The following result is a direct discrete-time analogue of Theorem 2.9.

**Theorem 2.10 (DT-LTI Representation Determines a DT-LTI System)** *Consider a discrete-time, LTI system representation  $(A, B, C, D)$ . This representation determines a discrete-time, linear time-invariant I/O system  $\mathcal{D}$ , specified by the following data:*

1. Time Set:  $\mathcal{T} = \mathbb{Z}$ .
2. Spaces:  $\mathcal{U} = \{u : \mathbb{Z} \rightarrow \mathbb{R}^m\}$ ,  $\mathcal{Y} = \{y : \mathbb{Z} \rightarrow \mathbb{R}^p\}$ , and  $\Sigma = \mathbb{R}^n$ .
3. State Transition Map: the state transition map is computed,

$$\varphi(k_1, k_0, x_0, u[\cdot]) = A^{k_1-k_0}x_0 + \sum_{j=k_0}^{k_1-1} A^{k_1-j-1}Bu[j]. \quad (2.160)$$

4. Readout Map: the readout map is computed,

$$r(k, x, u) = Cx + Du \quad (2.161)$$

5. I/O Map: The I/O map is computed,

$$\rho(k_1, k_0, x_0, u[\cdot]) = CA^{k_1-k_0}x_0 + C \left[ \sum_{j=k_0}^{k_1-1} A^{k_1-j-1}Bu[j] \right] + Du[k_1]. \quad (2.162)$$

**Proof** Follows directly from Proposition 2.13, Theorem 2.8, and verification of the time-invariance property.  $\square$

**Exercise 2.13** Provide the details of the proof of Theorem 2.10.

### 2.3.3 The Jordan Canonical Form

Let's take stock of where we're at in the study of LTI systems. Above, we showed that the state transition matrix of an LTI system is computed,

$$\Phi(t, t_0) = \exp(A(t - t_0)) \quad (\text{continuous-time}) \quad (2.163)$$

$$\Phi[k, k_0] = A^{k-k_0} \quad (\text{discrete-time}). \quad (2.164)$$

Although both formulas represent considerable simplifications over the time-varying case, there's still work to be done. In order to understand how linear, time-invariant systems evolve, we must actually be able to compute the matrix exponential and compute arbitrary powers of  $A$ . This ability will be essential when studying the stability of linear, time-invariant system, where we'll require explicit bounds on the size of the matrix exponential and the size of the matrix power.



Let's begin by studying the discrete-time state transition matrix. We know that, for an arbitrary matrix  $A \in \mathbb{C}^{n \times n}$ , there's no easy way to directly compute a power  $A^k$  for any given  $k$ . Instead of directly computing a power of  $A$ , what we might like to do is to compute a power of a *transformed* version of  $A$ , where the transformed version is easier to work with. The following lemma suggests that matrix transformations behave well under exponents.

**Lemma 2.4 (Similarity Transform and Matrix Power)** *Let  $A \in \mathbb{C}^{n \times n}$ . For any invertible matrix  $T \in \mathbb{C}^{n \times n}$  and any  $k \in \mathbb{Z}_{\geq 0}$ ,  $(T^{-1}AT)^k = T^{-1}A^kT$ .*

*Remark 2.25* A transformation of the form  $T^{-1}AT$  is called a *similarity transform* of  $A$ . If there exists a  $T$  for which  $T^{-1}AT = B$ , then  $A$  and  $B$  are said to be *similar matrices*. One may show that similar matrices share the same characteristic polynomial and eigenvalues.

**Proof** We'll prove this by induction on  $k$ . For the base case,  $k = 0$ , one has,  $(T^{-1}AT)^0 = I$  and  $T^{-1}A^0T = T^{-1}T = I$ . Thus, the base case holds. Now, assume for induction that the result is true for  $k > 0$ . One has,

$$(T^{-1}AT)^{k+1} = (T^{-1}AT)^k T^{-1}AT = T^{-1}A^k T T^{-1}AT = T^{-1}A^{k+1}T. \quad (2.165)$$

Thus, by induction on  $k$ , the proposed result holds.  $\square$

Lemma 2.4 tells us that, if we can transform a matrix  $A$  into a form whose exponents are easy to compute, we can easily recover the exponents of  $A$ . We illustrate this with the case of a diagonal matrix.

**Proposition 2.14 (Power of a Diagonalizable Matrix)** *Suppose  $A \in \mathbb{C}^{n \times n}$  is diagonalizable. That is, there exists an invertible matrix  $T \in \mathbb{C}^{n \times n}$  for which  $T^{-1}AT = D = \text{diag}(\lambda_1, \dots, \lambda_n)$ . Then, for any  $k \in \mathbb{Z}_{\geq 0}$ ,  $A^k$  is computed,*

$$A^k = T \text{diag}(\lambda_1^k, \dots, \lambda_n^k) T^{-1}. \quad (2.166)$$

**Proof** First, one may verify through an induction argument that, for a diagonal matrix  $D = \text{diag}(\lambda_1, \dots, \lambda_n)$ ,  $D^k = \text{diag}(\lambda_1^k, \dots, \lambda_n^k)$  for any  $k \in \mathbb{Z}_{\geq 0}$ . If there exists an invertible matrix  $T$  for which  $T^{-1}AT = D$ , one has that  $A = TDT^{-1}$ . Applying Lemma 2.4 and the formula for the exponential of a diagonal matrix, it follows that, for  $k \in \mathbb{Z}_{\geq 0}$ ,

$$A^k = T D^k T^{-1} = T \text{diag}(\lambda_1^k, \dots, \lambda_n^k) T^{-1}, \quad (2.167)$$

which is the desired result.  $\square$

We conclude that any nonnegative exponent of a diagonalizable matrix is easy to compute. To find any (nonnegative) exponent of a diagonalizable matrix  $A$ , all we need is the matrix  $T$  which diagonalizes  $A$ , the inverse of  $T$ , and the power of the diagonalization of  $A$  (which is easy to compute). This sketches out a basic technique—*transform and compute*—for calculating the state transition matrix in the discrete-time case. Now, we study the continuous-time case.

How might we calculate  $\exp(At)$  for a given matrix  $A$ ? An initial guess is that  $[\exp(At)]_{ij} = \exp([A]_{ij}t)$ , i.e. that taking the exponential of a matrix is the *same thing* as taking a term-by-term exponential of the entries of the matrix. However, a simple example shows that this is certainly *not* the case.

**Exercise 2.14 (Matrix Exponential is not Element-Wise)** Show that for the  $2 \times 2$  identity matrix,  $I_2$ ,

$$\exp \left( \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} t \right) \neq \begin{bmatrix} e^t & e^0 \\ e^0 & e^t \end{bmatrix}. \quad (2.168)$$

Now that we've ruled out the only "obvious" guess for calculating the matrix exponential, we must come up with some more clever techniques for computing the exponential. Let's follow a similar process to the discrete-time case—*transform and compute*. The first step in doing this is to verify that the matrix exponential "behaves well" under transformations.

**Proposition 2.15 (Similarity Transform and Matrix Exponential)** *Consider a matrix  $A \in \mathbb{C}^{n \times n}$ . For any invertible matrix  $T \in \mathbb{C}^{n \times n}$ ,*

$$\exp(T^{-1}AT) = T^{-1} \exp(A)T. \quad (2.169)$$

**Proof** Recall from Lemma 2.4 that, for all  $k \in \mathbb{Z}_{\geq 0}$ ,  $(T^{-1}AT)^k = T^{-1}A^kT$ . Applying this result to the partial sums of the exponential, we have,

$$\sum_{k=0}^p \frac{(T^{-1}AT)^k}{k!} = \sum_{k=0}^p \frac{T^{-1}A^kT}{k!} = T^{-1} \left( \sum_{k=0}^p \frac{A^k}{k!} \right) T. \quad (2.170)$$

Since matrix multiplication is a continuous operation, we can pass to the limit as  $p \rightarrow \infty$ . This yields,

$$\exp(T^{-1}AT) = \sum_{k=0}^{\infty} \frac{(T^{-1}AT)^k}{k!} = T^{-1} \left( \sum_{k=0}^{\infty} \frac{A^k}{k!} \right) T = T^{-1} \exp(A)T, \quad (2.171)$$

which completes the proof.  $\square$

Now that we've confirmed that the matrix exponential "behaves well" under transformations, we can find a transformation under which the exponential is easy to compute. As with the case of a matrix power, we'll find that the exponential of a diagonalizable matrix is easy to compute.

**Proposition 2.16 (Exponential of a Diagonalizable Matrix)** *Consider a diagonalizable matrix  $A \in \mathbb{C}^{n \times n}$  and an invertible transformation  $T \in \mathbb{C}^{n \times n}$  for which  $T^{-1}AT = \text{diag}(\lambda_1, \dots, \lambda_n)$ . The exponential of such a matrix may be computed,*

$$\exp(At) = T \text{diag}(e^{\lambda_1 t}, \dots, e^{\lambda_n t}) T^{-1}, \quad \forall t \in \mathbb{R}. \quad (2.172)$$

**Proof** First, we compute the exponential of a diagonal matrix,  $D = \text{diag}(\lambda_1, \dots, \lambda_n)$ . Since  $D^k = \text{diag}(\lambda_1^k, \dots, \lambda_n^k)$ , one has that,

$$\exp(Dt) = \sum_{k=0}^{\infty} \frac{(Dt)^k}{k!} = \text{diag} \left( \sum_{k=0}^{\infty} \frac{(\lambda_1 t)^k}{k!}, \dots, \sum_{k=0}^{\infty} \frac{(\lambda_n t)^k}{k!} \right) = \text{diag}(e^{\lambda_1 t}, \dots, e^{\lambda_n t}). \quad (2.173)$$

Now, suppose that  $T^{-1}AT = D = \text{diag}(\lambda_1, \dots, \lambda_n)$ . Then, applying Proposition 2.15,

$$\exp(At) = \exp(TDT^{-1}t) = T \exp(Dt) T^{-1} = T \text{diag}(e^{\lambda_1 t}, \dots, e^{\lambda_n t}) T^{-1}. \quad (2.174)$$

This completes the proof.  $\square$

These results tell us that, as long as the  $A$  matrix is diagonalizable, we can compute the state transition matrix in closed form. But, the general problem persists—not every matrix is diagonalizable. For instance, one may show that the matrix,

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \quad (2.175)$$

is non-diagonalizable. In order to determine a method of calculating the matrix exponential of *any* complex matrix  $A \in \mathbb{C}^{n \times n}$ , we'll study a (quite fragile) tool called the *Jordan canonical form* (or Jordan form, for short). This is a powerful theoretical tool that enables us to find decompositions of non-diagonalizable matrices that are amenable to computing exponents and exponentials.

Before we jump into the theory of the Jordan form, we must make a disclaimer: the Jordan form is an extremely useful tool for *theory* but is extremely impractical for *computation*. Later, we'll illustrate that the Jordan form of a matrix is extremely fragile to numerical perturbations, something that makes it generally unsuitable for computational use.

Let's motivate the Jordan form by examining what worked well about diagonalization in context of the state transition matrix. The key property of diagonalization for computing the state transition matrix is that a diagonal matrix has a *predictable structure* under exponents. That is, a diagonal matrix enjoys the property,

$$\begin{bmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_n \end{bmatrix}^k = \begin{bmatrix} \lambda_1^k & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_n^k \end{bmatrix}, \quad k \in \mathbb{Z}_{\geq 0}. \quad (2.176)$$

The main problem with diagonalization is that diagonalizability is *too strong* a condition—it's not always possible to transform a matrix  $A$  into a diagonal matrix via a similarity transform  $D = T^{-1}AT$ . What we'd like is a *relaxation* of diagonalizability that applies to any matrix and retains a predictable structure under exponentiation.

Consider the following idea: instead of enforcing a strictly diagonal structure, what if we enforce a *block-diagonal* structure? We recall the definition of a block diagonal matrix.

**Definition 2.23 (Block Diagonal Matrix)** A matrix  $D \in \mathbb{C}^{n \times n}$  is said to be block diagonal if there exist matrices  $A_i \in \mathbb{C}^{n_i \times n_i}$ ,  $i = 1, \dots, p$ , with  $\sum_{i=1}^p n_i = n$  and

$$D = \begin{bmatrix} A_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & A_p \end{bmatrix}, \quad (2.177)$$

where each 0 is an appropriately sized zero matrix. We write  $D = \text{blkdiag}(A_1, \dots, A_p)$ .

Since a block diagonal matrix has a similar structure to a diagonal matrix, it enjoys a similar behavior under exponentiation.

**Proposition 2.17 (Exponents of Block Diagonal Matrices)** Let  $D = \text{blkdiag}(A_1, \dots, A_p)$  be a block diagonal matrix. For any  $k \in \mathbb{Z}_{\geq 0}$ ,  $D^k = \text{blkdiag}(A_1^k, \dots, A_p^k)$ ,

$$\begin{bmatrix} A_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & A_p \end{bmatrix}^k = \begin{bmatrix} A_1^k & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & A_p^k \end{bmatrix}. \quad (2.178)$$

**Exercise 2.15** Verify the claim of Proposition 2.17.

Thus, we observe that—just like a diagonal matrix—a block diagonal matrix has a predictable structure under exponentiation. But, examining the exponent of a block diagonal matrix, we see that it’s not “good enough” to just have the block diagonal structure! In order to compute an exponent of a block diagonal matrix, we still need to compute the exponents of the individual blocks contained within the matrix. So, in addition to asking for a block-diagonal structure, we need the blocks themselves to have a “nice” structure for exponentiation. By relaxing a diagonal structure to an “almost-diagonal” structure, the following definition introduces exactly the type of block we will need.

**Definition 2.24 (Jordan Block)** Let  $\lambda \in \mathbb{C}$ . The Jordan block of size  $n$  corresponding to  $\lambda$  is the matrix  $J = \lambda I + N_0$ , where  $I$  is the  $n \times n$  identity matrix and  $N_0$  is an  $n \times n$  matrix of ones in the superdiagonal and zeros elsewhere. That is,  $J$  is the matrix in which all diagonal entries equal  $\lambda$  and entries just above the diagonal equal 1,

$$J = \begin{bmatrix} \lambda & 1 & 0 & \dots & 0 \\ 0 & \lambda & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda & 1 \\ 0 & 0 & \dots & 0 & \lambda \end{bmatrix}, \quad N_0 = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & \dots & 0 & 0 \end{bmatrix} \in \mathbb{C}^{n \times n} \quad (2.179)$$

The structure of a Jordan block as a matrix  $J = \lambda I + N_0$  makes its exponents particularly easy to compute. We know how to compute any exponent of  $\lambda I$ , since  $\lambda I$  is diagonal. Now, we show how to compute any exponent of  $N_0$ .

**Lemma 2.5 (Exponents of  $N_0$ )** Consider the  $n \times n$  Jordan block with respect to 0,  $N_0 \in \mathbb{C}^{n \times n}$ . The exponents of  $N_0$  respect the sequence,

$$N_0 = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & \dots & 0 & 0 \end{bmatrix}, \quad N_0^2 = \begin{bmatrix} 0 & 0 & 1 & \dots & 0 \\ & \ddots & \ddots & \ddots & \vdots \\ & & 0 & 0 & 1 \\ & & & 0 & 0 \\ 0 & & & & 0 \end{bmatrix}, \dots, \quad N_0^n = 0_{n \times n}. \quad (2.180)$$

That is, for each successive exponent, the superdiagonal of ones shifts one step to the right.

**Remark 2.26** A square matrix  $N$  for which  $N^k \neq 0$  for  $0 \leq k < n$  and  $N^k = 0$  for  $k \geq n$  is said to be a *nilpotent matrix* of order  $n$ .  $N_0$ , defined above, is an example of such a matrix.

**Exercise 2.16** Prove Lemma 2.5.

For  $N_0 \in \mathbb{C}^{n \times n}$ , we can use the “shift to the right” rule to compute  $N_0^k$  when  $k < n$ , and  $N_0^k = 0$  for  $k \geq n$ . Since for any Jordan block  $J$ ,  $J = \lambda I + N_0$ , this fact lets us compute an arbitrary exponent of *any* Jordan block in terms of the (known) exponents of  $N_0$ .

**Proposition 2.18 (Exponents of a Jordan Block)** Consider a Jordan block  $J \in \mathbb{C}^{n \times n}$  with respect to  $\lambda$ ,  $J = \lambda I + N_0$ . Then, for all  $k \in \mathbb{Z}_{\geq 0}$ ,

$$J^k = \sum_{j=0}^k \binom{k}{j} \lambda^{k-j} N_0^j. \quad (2.181)$$

**Proof** This proof follows from simple application of the binomial theorem, which states, for  $k \in \mathbb{Z}_{\geq 0}$  and  $A, B \in \mathbb{C}^{n \times n}$ ,

$$(A + B)^k = \sum_{j=0}^k \binom{k}{j} A^{k-j} B^j. \quad (2.182)$$

Let's apply this formula for  $A = \lambda I$  and  $B = N_0$ . We have that,

$$(\lambda I + N_0)^k = \sum_{j=0}^k \binom{k}{j} (\lambda I)^{k-j} N_0^j = \sum_{j=0}^k \binom{k}{j} \lambda^{k-j} N_0^j. \quad (2.183)$$

This is the desired formula.  $\square$

**Exercise 2.17** Expand the binomial formula to show that for a Jordan block  $J = \lambda I + N_0$ ,

$$J^k = \begin{bmatrix} \lambda^k \binom{k}{1} \lambda^{k-1} \binom{k}{2} \lambda^{k-2} & \dots & & \\ 0 & \lambda^k & \binom{k}{1} \lambda^{k-1} & \dots \\ \vdots & \vdots & \ddots & \ddots \\ 0 & 0 & \dots & \lambda^k \binom{k}{1} \lambda^{k-1} \\ 0 & 0 & \dots & 0 & \lambda^k \end{bmatrix}. \quad (2.184)$$

Let's summarize what we've found so far. We established that requiring *diagonalizability* is too strong an ask, and that *block diagonalizability* might be a suitable relaxation for computing exponents. Then, we showed that when each block has the structure of a Jordan block, we can compute its exponent. This means that for any block-diagonal matrix,

$$\begin{bmatrix} J_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & J_p \end{bmatrix}, \quad (2.185)$$

in which  $J_i$  is a Jordan block with respect to some constant  $\lambda_i$ , we can compute *any* power of the matrix. Thus, we've constructed a non-diagonal matrix form that we can easily exponentiate. This matrix form has the following special name.

**Definition 2.25 (Jordan Canonical Form)** Consider a matrix  $J \in \mathbb{C}^{n \times n}$ .  $J$  is said to be in Jordan canonical form (JCF) if there exist Jordan blocks  $J_i \in \mathbb{C}^{n_i \times n_i}$ ,  $i = 1, \dots, p$  such that  $\sum_{i=1}^p n_i = n$  and  $J = \text{blkdiag}(J_1, \dots, J_p)$ :

$$J = \begin{bmatrix} J_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & J_p \end{bmatrix} \in \mathbb{C}^{n \times n}, \quad J_i = \begin{bmatrix} \lambda_i & 1 & 0 & \dots & 0 \\ 0 & \lambda_i & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_i & 1 \\ 0 & 0 & \dots & 0 & \lambda_i \end{bmatrix} \in \mathbb{C}^{n_i \times n_i}. \quad (2.186)$$

Now that we've established a definition for the Jordan canonical form, we ask the essential question—can we transform any matrix into Jordan canonical form? That is, for any given matrix  $A \in \mathbb{C}^{n \times n}$ , does there exist an invertible matrix  $T \in \mathbb{C}^{n \times n}$  for which  $J = T^{-1}AT$  is in Jordan canonical form? Amazingly, the answer to this question is *yes*.

How might we construct the transformation into Jordan form? To start off, let's review how a transformation of a diagonalizable matrix  $A \in \mathbb{C}^{n \times n}$  into a diagonal matrix  $D \in \mathbb{C}^{n \times n}$  is constructed. In order to diagonalize a matrix  $A$ , one finds a linearly independent *basis of eigenvectors* of  $A$ —a linearly independent collection  $\{v_1, \dots, v_n\} \subseteq \mathbb{C}^{n \times n}$  for which  $Av_i = \lambda_i v_i$ , for some  $\lambda_i \in \mathbb{C}$ . Then, for such a collection, one has

$$\begin{bmatrix} | & & | \\ v_1 & \dots & v_n \\ | & & | \end{bmatrix}^{-1} A \begin{bmatrix} | & & | \\ v_1 & \dots & v_n \\ | & & | \end{bmatrix} = \begin{bmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_n \end{bmatrix}. \quad (2.187)$$

Thus, in the event that  $A$  is diagonalizable, one may use the eigenvectors of  $A$  to construct a matrix  $T$  for which  $T^{-1}AT$  is diagonal. We conclude that, in the event where  $A$  is *not* diagonalizable, we cannot form a basis  $\{v_1, \dots, v_n\}$  for  $\mathbb{C}^n$  of eigenvectors of  $A$ . Therefore, we have “too few” linearly independent eigenvectors in the non-diagonalizable case.

In order to construct a transformation of any matrix into a form *close* to a diagonal matrix—the Jordan canonical form—it seems reasonable that we might seek a *generalization* of the concept of an eigenvector. In order to define a *generalized eigenvector*, we start with the definition of an eigenvector. For  $A \in \mathbb{C}^{n \times n}$  and  $\lambda \in \mathbb{C}$ , we recall that a vector  $v \in \mathbb{C}^n$  is an eigenvector of  $A$  with eigenvalue  $\lambda$  if,

$$Av = \lambda v. \quad (2.188)$$

Rearranging this expression, we find that  $Av = \lambda v$  is equivalent to,

$$(A - \lambda I)v = 0. \quad (2.189)$$

An easy way to *generalize* the condition  $(A - \lambda I)v = 0$  is to require  $(A - \lambda I)^k v = 0$ , for some  $k \in \mathbb{N}$ . This leads to the following definition.

**Definition 2.26 (Generalized Eigenspace/Eigenvector)** Consider a matrix  $A \in \mathbb{C}^{n \times n}$  with eigenvalue  $\lambda \in \mathbb{C}$ . The *generalized eigenspace* of  $A$  corresponding to  $\lambda$  is the space,

$$K_\lambda(A) := \{v \in \mathbb{C}^n : (A - \lambda I)^m v = 0 \text{ for some } m \in \mathbb{N}\}. \quad (2.190)$$

Any nonzero vector  $v \in K_\lambda(A)$  is called a *generalized eigenvector* of  $A$  corresponding to the eigenvalue  $\lambda$ .

We quickly summarize a number of important properties of generalized eigenspaces.

**Proposition 2.19 (Properties of Generalized Eigenspaces)** Let  $A \in \mathbb{C}^{n \times n}$  be a matrix with eigenvalue  $\lambda \in \mathbb{C}$ . The generalized eigenspace  $K_\lambda(A)$  satisfies:

1. Eigenvector:  $K_\lambda(A)$  contains at least one eigenvector of  $A$ .
2. Subspace:  $K_\lambda(A)$  is a subspace.
3. Invariance:  $v \in K_\lambda(A)$  implies  $Av \in K_\lambda(A)$ .
4. Dimension: if the algebraic multiplicity<sup>4</sup> of  $\lambda$  is  $m$ , then  $\dim(K_\lambda(A)) = m$ .
5. Alternate Definition:  $K_\lambda(A) = \{v \in V : (T - \lambda I)^{\dim V} = 0\}$ .
6. Decomposition: For  $\lambda_1, \dots, \lambda_p \in \mathbb{C}$  the distinct eigenvalues of  $A$ ,  $\mathbb{C}^n$  is decomposed,

$$\mathbb{C}^n = K_{\lambda_1}(A) \oplus \dots \oplus K_{\lambda_n}(A), \quad (2.191)$$

where  $\oplus$  represents the direct sum<sup>5</sup> of subspaces.

**Proof** See Problem 2.20 and the starred, optional section below.  $\square$

The *generalized eigenvectors* of a matrix  $A$  form a candidate pool of vectors from which we will construct a transformation into Jordan form. Which generalized eigenvectors should we actually pick? To answer this question, we'll perform a cursory analysis of a transformation  $T$  which transforms  $A$  into a Jordan canonical form. Through this analysis, we'll determine necessary conditions on the generalized eigenvectors which enable a transformation into Jordan form. Suppose  $J = T^{-1}AT$  is a Jordan canonical form of  $A$ . Then, we must have,

$$TJ = AT \quad (2.192)$$

$$\begin{bmatrix} | & & | \\ V_{J_1} & \dots & V_{J_p} \\ | & & | \end{bmatrix} \begin{bmatrix} J_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & J_p \end{bmatrix} = A \begin{bmatrix} | & & | \\ V_{J_1} & \dots & V_{J_p} \\ | & & | \end{bmatrix}, \quad (2.193)$$

where each  $V_{J_i} = [v_{J_i}^{(1)}, \dots, v_{J_i}^{(n_i)}] \subseteq \mathbb{C}^{n \times n_i}$  is a block of vectors corresponding to the  $i$ 'th Jordan block,  $J_i$ . Let's focus on a single Jordan block, and see what necessary conditions this structure enforces. Examining the block  $J_i$ , we have

$$V_{J_i} J_i = A V_{J_i} \quad (2.194)$$

$$\begin{bmatrix} | & & | \\ v_{J_i}^{(1)} & \dots & v_{J_i}^{(n_i)} \\ | & & | \end{bmatrix} \begin{bmatrix} \lambda_i & 1 & 0 & \dots & 0 \\ 0 & \lambda_i & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_i & 1 \\ 0 & 0 & \dots & 0 & \lambda_i \end{bmatrix} = \begin{bmatrix} | & & | \\ A v_{J_i}^{(1)} & \dots & A v_{J_i}^{(n_i)} \\ | & & | \end{bmatrix}. \quad (2.195)$$

Now, let's zoom in on an individual column. For  $j = 1$ , we have that

$$A v_{J_i}^{(1)} = \lambda_i v_{J_i}^{(1)}, \quad (2.196)$$

implying that  $v_{J_i}^{(1)}$  must be an *eigenvector* of  $A$  with eigenvalue  $\lambda_i$ . For  $j > 1$ , we have,

<sup>4</sup> Recall that the algebraic multiplicity of an eigenvalue  $\lambda$  of  $A$  is the number of times it appears as a root of the characteristic polynomial of  $A$ . For instance, if a matrix  $A$  has the characteristic polynomial  $\chi_A(s) = (s - \lambda_1)^{m_1} \cdot \dots \cdot (s - \lambda_p)^{m_p}$ , the algebraic multiplicity of  $\lambda_1$  would be  $m_1$ .

<sup>5</sup> Let  $V_1, \dots, V_n$  be subspaces of a vector space  $V$ . One says that  $V$  is the *direct sum* of  $V_1, \dots, V_n$ , written  $V = V_1 \oplus \dots \oplus V_n$  if, for any  $v \in V$ , there exist unique  $v_i \in V_i$  for which  $v = v_1 + \dots + v_n$ .

$$\lambda_i v_{J_i}^{(j)} + v_{J_i}^{(j-1)} = A v_{J_i}^{(j)} \quad (2.197)$$

$$v_{J_i}^{(j-1)} = A v_{J_i}^{(j)} - \lambda_i v_{J_i}^{(j)} \quad (2.198)$$

$$v_{J_i}^{(j-1)} = (A - \lambda_i I) v_{J_i}^{(j)}. \quad (2.199)$$

This suggests that—if we zoom in on each Jordan block—the vectors associated with that block must satisfy a *recurrence* relation in which the first element of the recurrence is an eigenvector of  $A$  with eigenvalue  $\lambda_i$ . We define a sequence of vectors satisfying this recurrence as a *chain of generalized eigenvectors*.

**Definition 2.27 (Chain of Generalized Eigenvectors)** Let  $A \in \mathbb{C}^{n \times n}$  and  $\lambda \in \mathbb{C}$  be an eigenvalue of  $A$ . A sequence of vectors  $\{v^i\}_{i=1}^m \subseteq \mathbb{C}^n$  satisfying the properties,

1. Linear independence: the set  $\{v^i\}_{i=1}^m$  is linearly independent,
2. Recurrence: for  $(A - \lambda I)v^1 = 0$  and  $(A - \lambda I)v^i = v^{i-1}$  for  $i = 2, \dots, m$ ,

is called a *chain of generalized eigenvectors* of length  $m$ , corresponding to  $\lambda$ . A chain is said to be *maximal* if it cannot be extended while respecting (1) and (2).

Definition 2.27 suggests that a sequence of vectors satisfying properties (1) and (2) should be *generalized eigenvectors*. Let's confirm that this is the case before moving on.

**Lemma 2.6 (Chains are Composed of Generalized Eigenvectors)** Let  $A \in \mathbb{C}^{n \times n}$  and  $\lambda \in \mathbb{C}$  be an eigenvalue of  $A$ . Any chain of generalized eigenvectors corresponding to  $\lambda$  is composed of generalized eigenvectors belonging to  $K_\lambda(A)$ .

**Proof** Let  $\{v^i\}_{i=1}^m$  be a chain of generalized eigenvectors of  $A$ , corresponding to  $\lambda$ . We will prove that  $\{v^i\}_{i=1}^m \subseteq K_\lambda(A)$  by induction. We know that  $v^1$  is an eigenvector of  $A$  with eigenvalue  $\lambda$ , which implies  $v^1 \in K_\lambda(A)$ .

Suppose for induction that, for some  $i < m$ ,  $v^i$  is an element of  $K_\lambda(A)$ . We will show  $v^{i+1}$ , which satisfies  $(A - \lambda I)v^{i+1} = v^i$ , also belongs to  $K_\lambda(A)$ . Since  $v^i$  is a generalized eigenvector of  $A$ , there exists a  $k > 0$  for which  $(A - \lambda I)^k v^i = 0$ . Then, one has,

$$(A - \lambda I)v^{i+1} = v^i \quad (2.200)$$

$$(A - \lambda I)^{k+1}v^{i+1} = (A - \lambda I)^k v^i = 0. \quad (2.201)$$

We conclude that  $v^{i+1} \in K_\lambda(A)$ . Thus, each element of the chain belongs to  $K_\lambda(A)$ .  $\square$

We've now completed all of the setup required to construct a transformation into the Jordan form. We introduced *generalized eigenvectors*, which encompassed a wider class of vectors than (normal) eigenvectors, and identified special sets of generalized eigenvectors, termed *chains*, which were deemed necessary to construct a transformation into Jordan form.

In order to complete the Jordan form story, we must confirm that we can construct a basis for  $\mathbb{C}^n$  consisting of chains of generalized eigenvectors, and must prove that such a basis does indeed produce the desired transformation into Jordan canonical form.

**Theorem 2.11 (Transformation into Jordan Canonical Form)** Let  $A \in \mathbb{C}^{n \times n}$ . Suppose  $A$  has  $k$  linearly independent eigenvectors,  $v_1, \dots, v_k \in \mathbb{C}^n$ . Let  $\{v_j^i\}_{i=1}^{m_j}$  be the maximal Jordan chain with initial vector  $v_j^1 = v_j$ ,  $j = 1, \dots, k$ . Then, the set,

$$\beta = \{v_1^i\}_{i=1}^{m_1} \cup \dots \cup \{v_k^i\}_{i=1}^{m_k}, \quad (2.202)$$



forms a basis for  $\mathbb{C}^n$ . For  $T = [v_1^1, \dots, v_k^{m_k}] \in \mathbb{C}^{n \times n}$ , the matrix  $J = T^{-1}AT$  is in Jordan canonical form.

As the proof of this result is rather involved, we defer it to the starred, optional subsection below—the proof can be skipped without loss of continuity.

### 2.3.3.1 Computing the Jordan Canonical Form

Theorem 2.11 tells us that, for any matrix  $A \in \mathbb{C}^{n \times n}$ , we can compute a transformation  $T$  for which  $T^{-1}AT$  is in Jordan canonical form. What's more, the theorem statement provides a method of *constructing* this transformation. We summarize the proposed method of constructing the transformation into Jordan form with the following algorithm.

**Corollary 2.1 (Algorithm for Computing the Jordan Form)** *Let  $A \in \mathbb{C}^{n \times n}$ . The following procedure outlines a technique for computing a Jordan form of  $A$ .*

1. Eigenvalues: identify all distinct eigenvalues,  $\lambda_1, \dots, \lambda_p \in \mathbb{C}$  of  $A$ .
2. Eigenvectors: for each eigenvalue  $\lambda_i$ , find a maximal collection of linearly independent eigenvectors,  $v_{i,1}, \dots, v_{i,k_i}$  associated to  $\lambda_i$ .
3. Jordan chains: for each eigenvector  $v_{i,j}$  from step (2), compute the maximal Jordan chain starting at  $v_{i,j}$ . Repeat for all eigenvalues and eigenvectors.
4. Basis: construct an ordered basis for  $\mathbb{C}^n$  by taking the unions of all Jordan chains from step (3), preserving the order of each Jordan chain.
5. Matrix: construct a transformation matrix  $T$ , whose columns are the basis vectors from step (4). Be careful to preserve the order of the basis vectors!
6. Transformation: compute  $T^{-1}AT$  to find the Jordan form of  $A$ .

**Proof** Immediate from Theorem 2.11. □

We illustrate the use of this algorithm with a few examples.

*Example 2.2* Consider the matrix,

$$A = \begin{bmatrix} 3 & 0 & 0 \\ 1 & 3 & 0 \\ 0 & 0 & 3 \end{bmatrix}. \quad (2.203)$$

Let's work out the steps of the procedure.

1. Eigenvalues: it's apparent from the lower triangular structure of this matrix that the only eigenvalue of  $A$  is  $\lambda = 3$ .
2. Eigenvectors: next, we compute all linearly independent eigenvectors associated with the eigenvalue  $\lambda = 3$ . By inspection of the matrix, we find,

$$v_{3,1} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad v_{3,2} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}. \quad (2.204)$$

We cannot find any more linearly independent eigenvectors for  $\lambda = 3$ . Thus, we move on to step (3).

3. Jordan chains: now, we compute the maximal Jordan chains starting at  $v_{3,1}$  and  $v_{3,2}$ . Let's start with  $v_{3,1}$ . We wish to solve for  $v_{3,1}^2$  for which

$$(A - 3I)v_{3,1}^2 = v_{3,1} \quad (2.205)$$

$$\begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} v_{3,1}^2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}. \quad (2.206)$$

From this equality, it's clear that no such  $v_{3,1}^2$  exists. Thus, the Jordan chain starting at  $v_{3,1}$  is already maximal. Let's instead try the Jordan chain starting at  $v_{3,2}$ . We require,

$$(A - 3I)v_{3,2}^2 = v_{3,2} \quad (2.207)$$

$$\begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} v_{3,2}^2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}. \quad (2.208)$$

Selecting the vector,

$$v_{3,2}^2 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad (2.209)$$

accomplishes this.

4. Basis: now, we define a basis by taking the union of Jordan chains, preserving the order of each chain. We define,

$$\beta = \left\{ \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \right\}. \quad (2.210)$$

5. Matrix: now, we assemble the basis vectors into a transformation matrix, preserving the order of the basis. Define,

$$T = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}. \quad (2.211)$$

6. Transformation: finally, we compute  $T^{-1}AT$ . This yields,

$$T^{-1}AT = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 3 & 1 \\ 0 & 0 & 3 \end{bmatrix}, \quad (2.212)$$

which is in Jordan canonical form with two Jordan blocks,

$$J_1 = [3], J_2 = \begin{bmatrix} 3 & 1 \\ 0 & 3 \end{bmatrix}. \quad (2.213)$$

This example illustrates the following point—the Jordan form of a matrix is generally *not* unique—one can permute the order of the Jordan blocks and get a different Jordan form.

However, one may show that—*up to permutation of the block order*—the Jordan form of a matrix is unique. The interested reader is referred to [15] for proof of this fact.

This example also highlights a critical problem with the Jordan form, *numerical stability*. Suppose we're given the Jordan form matrix,

$$\begin{bmatrix} 3 & 0 & 0 \\ 0 & 3 & 1 \\ 0 & 0 & 3 \end{bmatrix}, \quad (2.214)$$

from the example above. Now, let's perturb one of the terms by arbitrarily small constant,  $\epsilon > 0$ , to get the matrix

$$\begin{bmatrix} 3 & 0 & 0 \\ 0 & 3 & 1 \\ 0 & 0 & 3 + \epsilon \end{bmatrix}. \quad (2.215)$$

Is this matrix still in Jordan form? No! Now, the lower block of the matrix is no longer a Jordan block, since  $3 + \epsilon \neq 3$  for  $\epsilon > 0$ . Thus, we find that for *any* arbitrarily small perturbation, the Jordan form of a matrix is at risk of changing structure. For instance, one may show that the Jordan form of the perturbed matrix above would be,

$$\begin{bmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 3 + \epsilon \end{bmatrix}. \quad (2.216)$$

Thus, we observe that the *structure* of the Jordan form is fragile to numerical perturbations. This illustrates the following general principle: the Jordan form is very useful for theoretical computation, but is impractical for numerical computation, where computer roundoff errors might arise. In summary:

*The JCF is generally good for theory and bad for application.*

### 2.3.3.2 Computing $\exp(A)$ and $A^k$

We finally have all the tools we need to meet the original goal we outlined at the beginning of the section: computing  $\exp(A)$  and  $A^k$  in closed form. Above, when motivating the Jordan form, we showed how to compute any exponent of a Jordan block. Now, we'll focus on computing the matrix exponential of a Jordan block.

There are a couple of ways of computing the matrix exponential of a Jordan block—we'll use a particularly slick method that takes advantage of *nilpotent matrices*, but more brute-force methods are possible. First, consider the following intermediate result.

**Lemma 2.7 (Commutativity & the Exponential)** *Let  $A, B \in \mathbb{C}^{n \times n}$ .*

$$AB = BA \implies \exp(A + B) = \exp(A) \exp(B). \quad (2.217)$$

**Proof** See Problem 2.13. □

This result mirrors the sum-product rule for the scalar exponential:  $e^{a+b} = e^a e^b$ . Since scalars always commute ( $ab = ba \ \forall a, b \in \mathbb{C}$ ) the scalar exponential rule actually follows from the lemma above! With this fact in mind, we state the following result.

**Proposition 2.20 (Exponential of a Jordan Block)** *Consider a Jordan block  $J = \lambda I + N_0 \in \mathbb{C}^{n \times n}$ . The exponential  $\exp(Jt)$  is computed,*

$$\exp(Jt) = e^{\lambda t} \begin{bmatrix} 1 & t & \dots & \frac{t^{n-1}}{(n-1)!} \\ 0 & 1 & \ddots & \vdots \\ \vdots & & \ddots & t \\ 0 & \dots & \dots & 1 \end{bmatrix} \in \mathbb{C}^{n \times n}. \quad (2.218)$$

**Proof** Consider a Jordan block  $J = \lambda I + N$ . We wish to compute  $\exp(Jt) = \exp((\lambda I + N)t)$ . We observe that the matrices  $\lambda I$  and  $N_0$  commute, since  $\lambda I N_0 = N_0 \lambda I$ . Applying Lemma 2.7, it follows that,

$$\exp(Jt) = \exp(\lambda I t + N_0 t) = \exp(\lambda I t) \exp(N_0 t). \quad (2.219)$$

Thus, in order to compute the exponential of  $J$ , we just need to compute the exponential of  $\lambda I$  and the exponential of  $N_0$ . Since  $\lambda I$  is diagonal, we know that  $\exp(\lambda I t) = e^{\lambda t} I$ . Now, we compute  $\exp(N_0 t)$ . We recall that  $N_0$  has the useful property that for every successive power of  $n$ , the diagonal of ones “shifts one to the right” and that  $N_0^n = 0$ . This means,

$$\exp(N_0 t) = I + N_0 t + \frac{1}{2!} N_0^2 t^2 + \frac{1}{3!} N_0^3 t^3 + \dots + \frac{1}{(n-1)!} N_0^{n-1} t^{n-1}, \quad (2.220)$$

since all exponents of degree  $n$  and higher vanish. All that remains is to compute the product  $e^{\lambda t} I \cdot \exp(N_0 t)$ . This gives,

$$\exp(Jt) = e^{\lambda t} \begin{bmatrix} 1 & t & \dots & \frac{t^{n-1}}{(n-1)!} \\ 0 & 1 & \ddots & \vdots \\ \vdots & & \ddots & t \\ 0 & \dots & \dots & 1 \end{bmatrix}, \quad (2.221)$$

which is the desired formula. □

Now that we’ve computed both  $J^k$  and  $\exp(Jt)$  for a Jordan block  $J$ , we state the general result concerning state transition matrices. As the following theorem simply collects results that we’ve already proven and puts them under the same umbrella, we provide no further proof.

**Theorem 2.12 (Jordan Forms for the LTI State Transition Matrices)** *Consider a matrix  $A \in \mathbb{C}^{n \times n}$ . Suppose  $T \in \mathbb{C}^{n \times n}$  is an invertible matrix for which  $J = T^{-1} A T$  is in the Jordan canonical form  $J = \text{blkdiag}(J_1, \dots, J_p)$ .*

1. *The continuous-time state transition matrix  $\Phi(t, t_0)$  with respect to  $A$  is computed,*

$$\Phi(t, t_0) = T \begin{bmatrix} e^{J_1(t-t_0)} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & e^{J_p(t-t_0)} \end{bmatrix} T^{-1}, \quad (2.222)$$

$$e^{J_i(t-t_0)} = e^{\lambda_i(t-t_0)} \begin{bmatrix} 1 & t & \dots & \frac{t^{n-1}}{(n-1)!} \\ 0 & 1 & \ddots & \vdots \\ \vdots & & \ddots & t \\ 0 & \dots & \dots & 1 \end{bmatrix}. \quad (2.223)$$

2. The discrete-time state transition matrix  $\Phi[k, k_0]$  with respect to  $A$  is computed,

$$\Phi[k, k_0] = T \begin{bmatrix} J_1^{k-k_0} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & J_p^{k-k_0} \end{bmatrix} T^{-1}, \quad (2.224)$$

$$J_i^{k-k_0} = \sum_{j=0}^{k-k_0} \binom{k-k_0}{j} \lambda_i^{k-k_0-j} N_0^j. \quad (2.225)$$

### 2.3.3.3 Constructing the Jordan Canonical Form ★

*This subsection is optional, and can be skipped without loss of continuity.*

In this subsection, we complete the proof of Theorem 2.11, and show that we can *always* construct a transformation into Jordan form.<sup>6</sup> Our proof will closely follow that of [5]. In order to keep our treatment sufficiently brief, we'll abstract away a few details—we'll write (check this) in places where we skip a few steps. Let's begin!

The first insight into the problem of proving Theorem 2.11 is that, instead of focusing on a general Jordan block, we should focus on a *nilpotent* block, a block  $N$  for which  $N^k = 0$  for some  $k > 0$ . We recall the following important result from linear algebra.

**Lemma 2.8 (Basis from Nilpotent Matrices)** *Let  $V$  be a finite dimensional vector space over  $\mathbb{C}$ . Let  $T : V \rightarrow V$  be a nilpotent linear transformation of order  $m$  - that is,  $T^m = 0$  and  $T^k \neq 0$  for  $k < m$ . There exists a basis of  $\mathbb{C}^n$  of the form,*

$$\{u_1, Tu_1, \dots, T^{a_1-1}u_1, u_2, \dots, T^{a_2-1}u_2, \dots, u_k, Tu_k, \dots, T^{a_k-1}u_k\}, \quad (2.226)$$

where each  $a_i > 0$  and  $T^{a_i}u_i = 0$  for  $1 \leq i \leq k$ .

**Proof** We'll give proof of a special case of this result, and refer the reader to [5], Proposition 8.45 for the details of the general case. Let  $T$  be a nilpotent of order  $m$ . Since  $T^{m-1} \neq 0$ , there exists a vector  $u \in V$  for which  $T^{m-1}u \neq 0$ . Now, consider the collection,

<sup>6</sup> This “always” comes with a couple of technical caveats. Here, by taking  $A \in \mathbb{C}^{n \times n}$ , we assume that the underlying field of the vector space is *complex*, and that complex transformations  $T \in \mathbb{C}^{n \times n}$  are allowed. It is *not* the case that for  $A \in \mathbb{R}^{n \times n}$ , there always exists a real transformation  $T \in \mathbb{R}^{n \times n}$  taking  $A$  into Jordan form (e.g. consider a real, diagonalizable matrix with complex eigenvalues). In the next chapter, we'll show how to construct a strictly real analogue of the Jordan form.

$$\{u, Tu, \dots, T^{m-1}u\}. \quad (2.227)$$

We will show that this is a linearly independent collection. Suppose there exist complex numbers  $c_0, \dots, c_{m-1}$  for which

$$c_0u + c_1Tu + \dots + c_{m-1}T^{m-1}u = 0. \quad (2.228)$$

Now, apply  $T^{m-1}$  to both sides. This yields,

$$c_0T^{m-1}u = 0 \Rightarrow c_0 = 0. \quad (2.229)$$

Next, apply  $T^{m-2}$  to both sides. This yields  $c_1 = 0$ . Repeating this process, we find that  $c_0 = \dots = c_{m-1} = 0$ , which implies the collection is linearly independent. In the event where  $\text{span}\{u, Tu, \dots, T^{m-1}u\} = V$ , this completes the proof. For the general case, we refer the reader to Proposition 8.45 of [5].  $\square$

By examining the structure of the basis proposed in Lemma 2.8, one notices a resemblance to the *chains* of generalized eigenvectors we defined above. Now, we show how to use Lemma 2.8 to generate a basis for  $\mathbb{C}^n$  composed of chains of generalized eigenvectors from each generalized eigenspace.

Note that, in the proof of the next result and in the results that follow, we will make use of an abstract linear transformation  $L_A : \mathbb{C}^n \rightarrow \mathbb{C}^n$ , defined  $L_A(v) = Av$  for all  $v \in \mathbb{C}^n$ . That is,  $L_A$  is the abstract linear transformation associated to left multiplication by  $A$ . Although this transformation may seem useless at a first glance, we'll find it useful when discussing the restriction of  $L_A$  to certain subspaces of  $\mathbb{C}^n$ . We'll see exactly how this shakes out in the following result.

**Lemma 2.9 (Basis for a Generalized Eigenspace)** *Let  $A \in \mathbb{C}^{n \times n}$  and  $K_\lambda(A)$  be a generalized eigenspace of  $A$ . Then, the following properties are satisfied:*

1. *Invariance:*  $K_\lambda(A)$  is invariant under  $L_A$ .
2. *Basis:*  $K_\lambda(A)$  has a basis of maximal chains of generalized eigenvectors of  $A$ .

**Proof** First, we will prove that  $K_\lambda(A)$  is invariant under  $L_A$ . That is, we will show  $v \in K_\lambda(A)$  implies  $Av \in K_\lambda(A)$ . Suppose  $v \in K_\lambda(A)$ . Then,  $v$  satisfies  $(A - \lambda I)^m v = 0$  for some  $m \in \mathbb{N}$ . This implies that,

$$(A - \lambda I)^{m+1}v = 0 \quad (2.230)$$

$$(A - \lambda I)^m(A - \lambda I)v = 0 \quad (2.231)$$

$$(A - \lambda I)^m(Av - \lambda v) = 0 \quad (2.232)$$

$$(A - \lambda I)^m Av = (A - \lambda I)^m \lambda v \quad (2.233)$$

$$(A - \lambda I)^m Av = 0. \quad (2.234)$$

We conclude that  $L_A(v) = Av \in K_\lambda(A)$  and that  $K_\lambda(A)$  is invariant under  $L_A$ . Using this fact, we can prove (2). If  $K_\lambda(A)$  is invariant under  $A$ , we can define a linear transformation  $T : K_\lambda(A) \rightarrow K_\lambda(A)$  by  $T = L_A|_{K_\lambda(A)}$ —the restriction of the linear transformation  $L_A$  to the subspace  $K_\lambda(A)$ . Since  $K_\lambda(A)$  is invariant under  $L_A$ , it makes sense that  $T$  is a transformation from  $K_\lambda(A) \rightarrow K_\lambda(A)$ .

Since  $T$  is the restriction of multiplication by  $A$  to  $K_\lambda(A)$ , it follows that the only eigenvalue of  $T$  is  $\lambda$ . Therefore, the transformation  $(T - \lambda I)$  has all zero eigenvalues and is nilpotent

(see the exercise below for a proof of this implication). By Lemma 2.8, we conclude there exists a basis of  $K_\lambda(A)$  consisting of vectors,

$$\{u_1, (T - \lambda I)u_1, \dots, (T - \lambda I)^{a_1-1}u_1, \dots, u_k, (T - \lambda I)u_k, \dots, (T - \lambda I)^{a_k-1}u_k\}. \quad (2.235)$$

Each sequence of vectors,  $u_i, (T - \lambda I)u_i, \dots, (T - \lambda I)^{a_i-1}u_i$  contained in this set is actually a maximal chain of generalized eigenvectors, written in reverse! We see this by applying the definition of  $T$  as left multiplication by  $A$ . We conclude that  $K_\lambda(A)$  has a basis consisting of maximal chains of generalized eigenvectors of  $A$ .  $\square$

**Exercise 2.18** Prove that  $T \in \mathcal{L}(V, V)$  is nilpotent if and only if its only eigenvalue is zero. *Hint: one direction has a simple proof using the Cayley-Hamilton theorem, which we'll discuss later in the course.*

Great! We've now shown that we can find a basis for each  $K_\lambda(A)$  consisting of a maximal chain of generalized eigenvectors. Next, we'll show that our *entire space* has a basis of generalized eigenvectors.

**Lemma 2.10 ( $\mathbb{C}^n$  has a Basis of Generalized Eigenvectors)** *Let  $A \in \mathbb{C}^{n \times n}$ . Then,  $\mathbb{C}^n$  has a basis consisting of generalized eigenvectors of  $A$ .*

**Proof** We will prove this by induction on  $n$ , the dimension of  $\mathbb{C}^n$ . In the base case,  $n = 1$ , this is obvious, since every nonzero vector is an eigenvector of  $A$ . Suppose for induction that, for some  $n \geq 1$ ,  $\mathbb{C}^k$  has a basis consisting of generalized eigenvectors of  $A$  for all  $1 \leq k < n$ . Consider an arbitrary eigenvalue  $\lambda$  of  $A$ . One can show with a little algebra that,

$$\mathbb{C}^n = \ker(A - \lambda I)^n \oplus \text{range}(A - \lambda I)^n, \quad (2.236)$$

(check this if you don't believe me). Now, we have two cases to consider. First, suppose  $\ker(A - \lambda I)^n = \mathbb{C}^n$ . Then, every element of  $\mathbb{C}^n$  must be a generalized eigenvector of  $A$  - the result trivially follows in this case.

Let's examine the other case. Suppose  $\ker(A - \lambda I)^n \neq \mathbb{C}^n$ , which implies  $\text{range}(A - \lambda I)^n \neq \{0\}$ . Since  $\lambda$  is known to be an eigenvalue of  $A$ , we also have that  $\ker(A - \lambda I)^m \neq \{0\}$ . This leads us to conclude,

$$0 < \dim \text{range}(A - \lambda I)^n < n. \quad (2.237)$$

By the same reasoning as in Lemma 2.9,  $\text{range}(A - \lambda I)^n$  is invariant under  $L_A$ . Now, let  $S$  be the linear transformation  $S : \text{range}(A - \lambda I)^n \rightarrow \text{range}(A - \lambda I)^n$  defined  $S = L_A|_{\text{range}(A - \lambda I)^n}$ —as the restriction of left multiplication by  $A$  to  $\text{range}(A - \lambda I)^n$ . By induction, there exists a basis of  $\text{range}(A - \lambda I)^n$  consisting of generalized eigenvectors of  $S$  (which are also generalized eigenvectors of  $A$  by definition of  $S$ ). Take this basis and append a basis of  $\ker(A - \lambda I)^n$ . This yields a basis of  $\mathbb{C}^n$  consisting of generalized eigenvectors of  $A$ . This completes the inductive step, and the lemma follows.  $\square$

Next, we'll show that we can decompose our space into a direct sum of generalized eigenspaces. First, we recall that  $V = V_1 \oplus \dots \oplus V_m$  (where the  $V_i$  are subspaces of  $V$ ) if all  $v \in V$  admit a unique decomposition  $v = v_1 + \dots + v_m$ ,  $v_i \in V_i$ .

**Lemma 2.11 (Generalized Eigenspace Decomposition)** *Let  $A \in \mathbb{C}^{n \times n}$  and let  $\lambda_1, \dots, \lambda_p$  be the distinct eigenvalues of  $A$ . Then,  $\mathbb{C}^n$  decomposes as the direct sum,*

$$\mathbb{C}^n = K_{\lambda_1}(A) \oplus \dots \oplus K_{\lambda_p}(A). \quad (2.238)$$

**Proof** First, we show that each generalized eigenvector corresponds to only one generalized eigenvalue of  $A$ . Suppose  $v$  corresponds to both  $\lambda$  and  $\mu$ , and that  $m$  is the smallest integer for which  $(A - \mu I)^m v = 0$ . Then, since  $m \leq n$ ,

$$0 = (A - \lambda I)^n v \quad (2.239)$$

$$= ((A - \mu I) + (\mu - \lambda)I)^n v \quad (2.240)$$

$$= \sum_{k=0}^n \binom{n}{k} (\mu - \lambda)^{n-k} (A - \mu I)^k v. \quad (2.241)$$

Applying  $(A - \mu I)^{m-1}$  to both sides yields  $0 = (\mu - \lambda)^n (A - \mu I)^{m-1} v$ . Since  $(A - \mu I)^{m-1} v \neq 0$ , we conclude  $\mu = \lambda$  and that  $v$  must correspond to *one* eigenvalue. Next, we will show that any collection of generalized eigenvectors corresponding to different generalized eigenspaces,

$$\{v_1, \dots, v_p\}, \quad v_i \in K_{\lambda_i}(A), \quad \lambda_i \neq \lambda_j, \quad (2.242)$$

is linearly independent. Suppose for contradiction that there exists a linearly dependent collection,  $\{v_1, \dots, v_p\}$ . Then, there exist constants  $a_1, \dots, a_p$  - not all zero - for which  $a_1 v_1 + \dots + a_p v_p = 0$ . Apply  $(A - \lambda_p I)^n$  to both sides to get,

$$a_1 (A - \lambda_p I)^n v_1 + \dots + a_p (A - \lambda_p I)^n v_p = 0 \quad (2.243)$$

$$a_1 (A - \lambda_p I)^n v_1 + \dots + a_{p-1} (A - \lambda_p I)^n v_{p-1} = 0, \quad (2.244)$$

since  $v_p$  is a generalized eigenvector of  $\lambda_p$  and therefore satisfies  $(A - \lambda_p I)^n v_p = 0$ . Each remaining term *must* be nonzero, otherwise each remaining  $v_i$  would be a generalized eigenvector corresponding to two eigenvalues, which we proved above cannot happen. Yet, we know that, for  $k \neq p$ ,

$$(A - \lambda_k I)^n (A - \lambda_p I)^n v_k = (A - \lambda_p I)^n (A - \lambda_k I)^n v_k = 0. \quad (2.245)$$

So,  $(A - \lambda_k I)^n v_k$  is a generalized eigenvector corresponding to  $\lambda_p$ . This contradicts what we proved above! We conclude the collection is linearly independent.

Finally, we show the direct sum property. By Lemma 2.10, it follows that  $\mathbb{C}^n = K_{\lambda_1}(A) + \dots + K_{\lambda_p}(A)$ . Now, we need to show that this sum is a *direct sum*. It is a nice fact from linear algebra that  $V = V_1 \oplus \dots \oplus V_p$  if  $V = V_1 + \dots + V_p$  and  $v_1 + \dots + v_p = 0$ ,  $v_i \in V_i$  implies  $v_i = 0$  for all  $i$ . Suppose  $v_1 + \dots + v_p = 0$ , for  $v_i \in K_{\lambda_i}(A)$ . Since generalized eigenvectors corresponding to distinct eigenvalues are linearly independent, each  $v_i$  must be zero. We conclude that  $\mathbb{C}^n = K_{\lambda_1}(A) \oplus \dots \oplus K_{\lambda_p}(A)$ , which completes the proof.  $\square$

We're almost there! We've shown that  $\mathbb{C}^n$  admits a decomposition into a direct sum of generalized eigenspaces *and* that each generalized eigenspace has a basis of chains of generalized eigenvectors of  $A$ . All that remains is to put these facts together to conclude Theorem 2.11.

**Proof (Of Theorem 2.11 on Existence of a Transformation into Jordan Form)**

Let  $A \in \mathbb{C}^{n \times n}$  and let  $\lambda_1, \dots, \lambda_p$  be the distinct eigenvalues of  $A$ . Then, by Lemma 2.11,  $\mathbb{C}^n = K_{\lambda_1}(A) \oplus \dots \oplus K_{\lambda_p}(A)$ . As such, for any bases  $\beta_1, \dots, \beta_p$  of each of the generalized eigenspaces, the transformation  $[A]_{\beta_1 \cup \dots \cup \beta_p} = T^{-1} A T$ ,  $T = [\beta_1, \dots, \beta_p]$  should be in block diagonal form. Pick each basis  $\beta_i$  to be a basis for  $K_{\lambda_i}(A)$ , consisting of maximal chains of generalized eigenvectors (each starting at a true eigenvector of  $\lambda_i$ ), as in Lemma 2.9. Each resulting block is a Jordan block, and the resulting matrix is in Jordan canonical form.  $\square$



Phew, what a proof! We note that this is *not* the only proof of the existence of the Jordan form - rather, this is one that generalizes well to arbitrary linear operators on finite dimensional vector spaces beyond  $\mathbb{C}^n$ . Other, shorter proofs that are more computation-heavy are possible. The interested reader is directed to [18] for the details of one such proof.

### 2.3.4 Further Reading

The treatment of the matrix exponential and state transition matrix in this section was primarily influenced by those of [25], [8], [14], and [27]. The proof of the construction of the Jordan form closely follows that of [5], and the algorithm for computing the Jordan form is from [19]. As mentioned above, for a more computational proof of the Jordan form theorem, we recommend the treatment of [18].

### 2.3.5 Problems

**Problem 2.12 (Zero-Order Hold Discretization)** In this problem, we'll show how a continuous-time linear can be *exactly* discretized into a discrete-time linear system. Consider the continuous-time system representation  $(A(\cdot), B(\cdot), C(\cdot), D(\cdot))$ ,

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) \quad (2.246)$$

$$y(t) = C(t)x(t) + D(t)u(t), \quad (2.247)$$

where  $x(t) \in \mathbb{R}^n$ ,  $u(t) \in \mathbb{R}^m$ , and  $y(t) \in \mathbb{R}^p$ . Consider a strictly increasing sequence of sampling times,  $\{t_k\}_{k \in \mathbb{Z}} \subseteq \mathbb{R}$ , satisfying  $t_k < t_{k+1}$  for all  $k \in \mathbb{Z}$ . Suppose the input signals  $u(\cdot)$  to the continuous-time system are constant on the sampling intervals. That is, for every input signal  $u(\cdot)$  to the continuous-time system, there exists a sequence  $u[\cdot] : \mathbb{Z} \rightarrow \mathbb{R}^m$  for which  $u(t) = u[k] \in \mathbb{R}$  for all  $t \in [t_k, t_{k+1})$ .

1. Show there exists a discrete-time system representation  $(\hat{A}[\cdot], \hat{B}[\cdot], \hat{C}[\cdot], \hat{D}[\cdot])$  such that for all initial conditions  $x_0 = x(t_0) = x[0] \in \mathbb{R}^n$ , solutions to the system,

$$x[k+1] = \hat{A}[k]x[k] + \hat{B}[k]u[k] \quad (2.248)$$

$$y[k] = \hat{C}[k]x[k] + \hat{D}[k]u[k], \quad (2.249)$$

satisfy  $x[k] = x(t_k)$  and  $y[k] = y(t_k)$  for all  $k \in \mathbb{Z}$ , where  $x(t_k)$  and  $y(t_k)$  are the state and output of the continuous-time system at time  $t_k$ . This tells us that we can *exactly* discretize the continuous-time LTV system.

2. Now, suppose each matrix in the continuous-time system representation is constant,

$$(A(\cdot), B(\cdot), C(\cdot), D(\cdot)) = (A, B, C, D). \quad (2.250)$$

Further, assume that for each  $k \in \mathbb{Z}$ ,  $t_{k+1} - t_k = \Delta$ . Using your answer to (1), show that there exists a discrete-time LTI system representation  $(\hat{A}, \hat{B}, \hat{C}, \hat{D})$  which exactly discretizes the continuous-time LTI system.

Discretization in which one holds an input signal constant across a sampling period is referred to as *zero-order hold (ZOH)* discretization.

**Problem 2.13 (Commutativity & The Exponential ★)** Let  $A, B \in \mathbb{R}^{n \times n}$ . In this problem, we'll prove that commutativity of  $A$  and  $B$ ,  $AB = BA$ , implies  $\exp(A + B) = \exp(A)\exp(B)$ . Notably, we'll give a proof that uses an *existence and uniqueness* argument, rather than a direct algebraic argument.

1. Give examples of matrices  $A, B$  for which  $\exp(A + B) \neq \exp(A)\exp(B)$ . *Feel free to use a computational tool to experiment with different matrices.*
2. Using an existence and uniqueness argument, prove that  $AB = BA$  implies  $\exp(A + B) = \exp(A)\exp(B)$ . *Hint: set up an initial value problem involving  $A$  and  $B$ .*

**Problem 2.14 (Algebraic Properties of the Matrix Exponential)** Let  $A \in \mathbb{R}^{n \times n}$ . Prove the following properties of the matrix exponential:

1. For every  $t_1, t_0 \in \mathbb{R}$ ,  $\exp(A(t_1 + t_0)) = \exp(At_1)\exp(At_0) = \exp(At_0)\exp(At_1)$ .
2. For an eigenvalue-vector pair  $(\lambda, v)$  of  $A$ ,  $(e^\lambda, v)$  is an eigenvalue-vector pair of  $\exp(A)$ .
3.  $\det(\exp A) = e^{\text{tr } A}$ .
4.  $(\exp(A))^{-1} = \exp(-A)$ .

*Hint: read the result of Problem 2.13 before attempting these problems.*

**Problem 2.15 (Some State Transition Matrices [10])** Sometimes, we can use the matrix exponential to gain insight into the structure of linear, time-varying initial value problems. Consider the following two problems.

1. Suppose  $A, B \in \mathbb{R}^{n \times n}$ . Show that the unique solution to the initial value problem,

$$\dot{x}(t) = e^{-At} B e^{At} x(t), \quad x(t_0) = x_0, \quad (2.251)$$

is given by  $x(t) = e^{-At} e^{(A+B)(t-t_0)} e^{At_0} x_0$ .

2. Let  $f : \mathbb{R} \rightarrow \mathbb{R}$  be a continuous function and  $A \in \mathbb{R}^{n \times n}$ . Show that the state transition matrix of the system,

$$\dot{x}(t) = f(t)Ax(t), \quad (2.252)$$

is computed  $\Phi(t, t_0) = \exp[\int_{t_0}^t f(\tau) d\tau A]$ .

**Problem 2.16 (The Complex Jordan Block)** Consider a matrix  $A \in \mathbb{R}^{2 \times 2}$ . If  $A$  has complex eigenvalues, there are sometimes more convenient transformations than the (complex) Jordan form! Suppose  $A$  has complex eigenvalues  $\sigma \pm j\omega$ .

1. Prove there exists a matrix  $T \in \mathbb{C}^{n \times n}$  for which,

$$T^{-1}AT = \begin{bmatrix} \sigma & \omega \\ -\omega & \sigma \end{bmatrix}. \quad (2.253)$$

2. Show that the exponential of  $T^{-1}AT$  is computed,

$$\exp(T^{-1}AT) = \begin{bmatrix} e^{\sigma t} \cos(\omega t) & e^{\sigma t} \sin(\omega t) \\ -e^{\sigma t} \sin(\omega t) & e^{\sigma t} \cos(\omega t) \end{bmatrix}. \quad (2.254)$$

**Problem 2.17 (The Floquet Decomposition ★)** Consider the system  $\dot{x}(t) = A(t)x(t)$ , in which  $A(\cdot)$  is periodic with period  $T > 0$ ,  $A(t + T) = A(t)$  for all  $t \in \mathbb{R}$ . The basic idea of the *Floquet decomposition* is that, by constructing a time-varying transformation that “syncs up” with the periodicity of  $A(\cdot)$ , we can use time-invariant tools to study a time-varying system.

1. Let  $\Phi(t, t_0)$  denote the state transition matrix of  $\dot{x}(t) = A(t)x(t)$ . Show that  $\Phi(t + T, 0) = \Phi(t, 0)\Phi(T, 0)$ .
2. Prove that for every nonsingular matrix  $B \in \mathbb{C}^{n \times n}$ , there exists a matrix  $A \in \mathbb{C}^{n \times n}$  for which  $\exp(A) = B$ . *Hint: the complex (scalar) logarithm is defined on the nonzero complex numbers.*
3. Prove there exists an  $R \in \mathbb{C}^{n \times n}$  for which  $\Phi(T, 0) = \exp(TR)$ .
4. Consider a time-varying transformation  $P : \mathbb{R} \rightarrow \mathbb{C}^{n \times n}$ , for which

$$P(t)^{-1} = \Phi(t, 0)e^{-tR}. \quad (2.255)$$

Show that  $P(t)$  is in fact invertible for all  $t$ . Then, prove that for any  $t, t_0 \in \mathbb{R}$ ,

$$\Phi(t, t_0) = P(t)^{-1}e^{R(t-t_0)}P(t_0). \quad (2.256)$$

Comment on the significance of this result. *The eigenvalues of the  $R$  matrix, called Floquet multipliers, help in determining the stability of periodic systems.*

**Problem 2.18 (Fun with Jordan Forms)** In the following problem, we’ll get some practice with Jordan forms.

1. Find a Jordan canonical form of each of the following matrices,

$$A_1 = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & -1 & 2 \end{bmatrix}. \quad (2.257)$$

2. Calculate the matrix exponentials  $\exp(A_i t)$ ,  $i = 1, 2$ .
3. For each of  $A_i$ , illustrate how an arbitrarily small numerical perturbation can change the structure of the Jordan form. How do you think this affects numerical computation of the Jordan form?

**Problem 2.19 (Schur Triangulation)** The Jordan normal form—though convenient for theoretical calculations—suffers from a number of numerical problems. Here, we consider an alternative technique for computing the matrix exponential, based on the *Schur triangulation* of a matrix.

1. A matrix  $T \in \mathbb{C}^{n \times n}$  is said to be *unitary* if  $TT^* = T^*T = I$ . Prove the *Schur triangulation theorem*, which states that, for any  $A \in \mathbb{C}^{n \times n}$ , there exists a unitary matrix  $T$  for which  $U := T^*AT$  is upper triangular. *Hints: Proceed by induction.*
2. Determine a method for computing the matrix exponential of an upper triangular matrix. *Your method does not have to be computationally efficient, it just needs to work.*
3. Comment on the benefits and drawbacks of using your Schur triangulation method versus the Jordan normal form method of calculating the matrix exponential.

If you’re interested in reading about more ways of computing the matrix exponential, check out the paper [Nineteen Dubious Ways to Compute the Exponential of a Matrix, Twenty-Five Years Later](#).

**Problem 2.20 (Some Invariant Subspace Theory ★)** Fundamentally, the construction of the Jordan canonical form relies on the fact that the generalized eigenspaces are *invariant subspaces* - subspaces  $V \subseteq \mathbb{C}^n$  satisfying  $AV \subseteq V$ . In this problem, we'll study some basic properties of invariant subspaces and see how they relate to the Jordan form.

1. Let  $T : V \rightarrow V$  be a linear transformation on an  $n$ -dimensional vector space  $V$  over  $\mathbb{K}$ . A subspace  $M \subseteq V$  is said to be  $T$ -invariant if  $Tx \in M$  for all  $x \in M$ . Suppose that  $V$  is the direct sum of two subspaces  $M_1, M_2 \subseteq V$ ,  $V = M_1 \oplus M_2$ . If both  $M_1$  and  $M_2$  are  $T$ -invariant, prove there exists a basis  $\beta$  for  $V$  in which  $T$  has the matrix representation,

$$[T]_\beta = A = \begin{bmatrix} A_{11} & 0 \\ 0 & A_{12} \end{bmatrix} \in \mathbb{K}^{n \times n}, \quad (2.258)$$

where  $\dim(M_1)$  and  $\dim(M_2)$  equal the sizes of  $A_{11}$  and  $A_{12}$ . Argue that the restrictions  $T|_{M_1} : M_1 \rightarrow M_1$  and  $T|_{M_2} : M_2 \rightarrow M_2$  are well-defined maps.

2. We define a generalized eigenspace of the linear transformation  $T$  to be a space,

$$K_\lambda(T) = \{v \in V : (T - \lambda I)^m = 0 \text{ for some } m \in \mathbb{N}\} \subseteq V, \quad (2.259)$$

where  $\lambda$  is an eigenvalue of  $T$ . Prove the following:

- a.  $K_\lambda(T)$  contains at least one eigenvector of  $T$ .
  - b.  $K_\lambda(T)$  is a subspace.
  - c.  $K_\lambda(T)$  is  $T$ -invariant.
  - d.  $K_\lambda(T) = \{v \in V : (T - \lambda I)^{\dim V} = 0\}$ .
3. Suppose  $T$  has two distinct eigenvalues,  $\lambda_1$  and  $\lambda_2$ , and  $V = K_{\lambda_1}(T) \oplus K_{\lambda_2}(T)$ . Suppose for each  $i = 1, 2$ , the sets

$$\beta_i = \{v_i, (T - \lambda_i I)v_i, \dots, (T - \lambda_i I)^{m_i-1}v_i\}, \quad (2.260)$$

form bases for  $K_{\lambda_1}$  and  $K_{\lambda_2}$ . Construct a basis  $\beta$  for  $V$  in which  $T$  is in Jordan canonical form.

## 2.4 Impulse Response & Transfer Functions

Thus far, our study of linear systems has focused primarily on the *state equations*,

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) \quad (2.261)$$

$$x[k+1] = A[k]x[k] + B[k]u[k], \quad (2.262)$$

of state space representations. We spent some time studying the existence and uniqueness of solutions to these equations, as well as establishing the structure of their solutions via the state transition matrix. At large, we haven't taken a closer look at the output equations.

In this section, this all changes! Instead of looking at the state equation of the system to gain insight into its behavior, we'll zoom out and examine how the input to a system directly influences its output. The border between these two approaches hints at two perspectives on control theory.

On one side, we have the *internal* (state space) approach, where we look into the internal dynamics of the system to make conclusions about its behavior. On the other side, we have the *input/output (I/O) approach*, where we focus directly on the relationship between the input and output. We'll continue to discuss the interplay between these approaches, as well as their various merits, as the course progresses.

This section is outlined as follows. First, we'll study discrete-time systems from the I/O perspective, and will focus on an object called the *impulse response*. Following this, we'll take on the more challenging task of defining the impulse response of a continuous-time system, taking a short detour to discuss the Dirac delta distribution along the way. We'll conclude with a discussion of two fundamental transforms—the Laplace and  $\mathcal{Z}$ -transforms—which take advantage of the unique structure of the impulse response to transform real, LTI systems into complex, algebraic functions. Let's begin!

### 2.4.1 Impulse Response of Discrete-Time Systems

We'll begin by studying discrete-time systems from the I/O perspective. Although we usually start with continuous-time and translate to discrete-time, here, we'll find taking the opposite approach to be helpful. In particular, we'll find that the discrete-time theory is considerably easier and provides substantial insight into the continuous-time case.

Let's recall what we know about the I/O behavior of a discrete-time system. Recall that, in Section 2, we stated that the input/output map  $\rho$  of a discrete-time linear, time-varying system representation is calculated,

$$\rho(k_1, k_0, x_0, u[\cdot]) = \underbrace{C[k_1]\Phi[k_1, k_0]x_0}_{\text{Zero-Input Response}} + \underbrace{C[k_1] \sum_{j=k_0}^{k_1-1} \Phi[k_1, j+1]B[j]u[j] + D[k_1]u[k_1]}_{\text{Zero-State Response}}, \quad (2.263)$$

where we can split up the I/O map into the zero-input and zero-state response terms. From this formula, we observe that the zero-state response *entirely* characterizes how inputs interact with the response of the system. Thus, if we'd like to develop the input/output perspective, we should focus on the behavior of the zero-state response.

Let's examine the formula for the zero-state response in greater detail. We know that the zero-state response is calculated,

$$\rho(k_1, k_0, 0, u[\cdot]) = C[k_1] \sum_{j=k_0}^{k_1-1} \Phi[k_1, j+1] B[j] u[j] + D[k_1] u[k_1], \quad (2.264)$$

We observe that the zero-state response at any given moment *only* depends on the history of inputs to the system—there is *no* presence of state in the zero-state response whatsoever! This hints at an interesting idea in the input/output approach to systems: assuming the initial state of the system is fixed (for example at  $x_0 = 0$ ), we can come up with a map from a pair of times and an input signal that maps *directly* to the value of an output signal. By studying the structure underlying such a map, perhaps we could come up with more elegant methods of studying the input/output behavior of a system than directly using the I/O map  $\rho$ . Amazingly, for linear systems, all we need to characterize this direct input-to-output function is the response of the system to a single input: the *unit impulse*.

**Definition 2.28 (Discrete-Time Unit Impulse)** The discrete-time unit impulse function is the function  $\delta[\cdot] : \mathbb{Z} \rightarrow \mathbb{R}$ , defined,

$$\delta[k] = \begin{cases} 1 & k = 0, \\ 0 & k \neq 0. \end{cases} \quad (2.265)$$

The shifted unit impulse,  $\delta^{k_0}[\cdot] : \mathbb{Z} \rightarrow \mathbb{R}$ , is defined  $\delta^{k_0}[k] = \delta[k - k_0]$ , for a fixed  $k_0 \in \mathbb{Z}$ .

Thus, the discrete-time unit impulse function (also called the *unit pulse* or *unit sample* function) is a scalar-valued discrete-time signal that jumps up to 1 at time  $k = 0$  and is zero elsewhere. We can *shift* the unit impulse to jump up at time  $k_0$  by defining  $\delta^{k_0}[k] = \delta[k - k_0]$ . Using the unit impulse function, we define the impulse response of a discrete-time linear system. We begin with the SISO case, and then generalize to the MIMO case.

**Definition 2.29 (Impulse Response of a DT-SISO System)** Consider a SISO, discrete-time linear I/O system with I/O map  $\rho$ . The impulse response map of the system is the map  $h[\cdot, \cdot] : \mathbb{Z} \times \mathbb{Z} \rightarrow \mathbb{R}$ , defined,

$$h[k, k_0] = \begin{cases} \rho(k, k_0, 0, \delta^{k_0}[\cdot]) & k \geq k_0 \\ 0 & k < k_0, \end{cases} \quad (2.266)$$

where  $\delta^{k_0}[k] = \delta[k - k_0]$ , the unit impulse applied at time  $k_0$ .

*Remark 2.27* As opposed to being defined on  $\mathbf{T} = \{(k_1, k_0) \in \mathbb{Z} \times \mathbb{Z} : k_1 \geq k_0\}$  (like the I/O map, which requires  $k_1 \geq k_0$ ), the impulse response map is defined on all of  $\mathbb{Z} \times \mathbb{Z}$ . This proves convenient in calculations.

Thus, we define the impulse response  $h[k, k_0]$  of a discrete-time linear system to be the zero-state response of a system to a unit impulse applied at time  $k_0$ . For all  $k < k_0$ , we define  $h[k, k_0]$  to be zero, as the I/O map is not defined on this domain. Defining the impulse response to be zero for  $k < k_0$  is justifiable, since we wouldn't expect a *causal* linear system to stray from zero if it has zero initial condition and zero input applied—causality implies that, for  $k < k_0$ , the system behavior is not affected by the impulse at time  $k_0$ . Now, we generalize this definition to a MIMO system.

**Definition 2.30 (Impulse Response of a MIMO System)** Consider a MIMO discrete-time, linear I/O system with I/O map  $\rho$ , input-value space  $\mathbb{R}^m$ , and output-value space  $\mathbb{R}^p$ . The impulse response map of the system is the map  $H[\cdot, \cdot] : \mathbb{Z} \times \mathbb{Z} \rightarrow \mathbb{R}^{p \times m}$ , defined

$$H[k, k_0] = \begin{bmatrix} h_{11}[k, k_0] & \dots & h_{1m}[k, k_0] \\ \vdots & \ddots & \vdots \\ h_{p1}[k, k_0] & \dots & h_{pm}[k, k_0] \end{bmatrix} \in \mathbb{R}^{p \times m} \quad (2.267)$$

$$h_{ij}[k, k_0] = \begin{cases} \rho(k, k_0, 0, \delta^{k_0}[\cdot]e_j)_i & k \geq k_0 \\ 0 & k < k_0, \end{cases} \quad (2.268)$$

where  $\rho(k, k_0, 0, \delta^{k_0}[\cdot]e_j)_i \in \mathbb{R}$  is the  $i$ 'th component of the zero-state response to an input  $\delta^{k_0}[k]e_j = (0, \dots, \delta[k - k_0], \dots, 0) \in \mathbb{R}^m$  containing a unit impulse function in its  $j$ 'th component.

In summary, the impulse response of a MIMO system is a matrix-valued function, in which entry  $ij$  is the response at time  $k$  of output coordinate  $i$  to an impulse applied in input coordinate  $j$  at time  $k_0$ . In order to ensure that the MIMO impulse response map is conceptually well-defined, it pays to check that each element  $h_{ij}$  corresponds to the impulse response map of a SISO system. The following exercise asks you to verify this.

**Exercise 2.19 (The MIMO Impulse Response is Well-Defined)** Consider a discrete-time, linear I/O system with input-value space  $\mathbb{R}^m$ , output-value space  $\mathbb{R}^p$ , and I/O map  $\rho$ . Prove there exists a SISO system with I/O map  $\rho_i^j$  satisfying  $\rho_i^j(k, k_0, 0, u_j[\cdot]) = \rho(k, k_0, 0, u_j[\cdot]e_j)_i$ , for all  $u_j[\cdot] : \mathbb{Z} \rightarrow \mathbb{R}$ .

Now that we've defined the impulse response of a linear I/O system, let's calculate the impulse response corresponding to a discrete-time, linear system representation.

**Proposition 2.21 (Impulse Response Map of a DT System Representation)** Consider a discrete-time, linear system representation  $(A[\cdot], B[\cdot], C[\cdot], D[\cdot])$ . The impulse response map of the system is computed,

$$H[k, k_0] = \begin{cases} C[k]\Phi[k, k_0 + 1]B[k_0] & k > k_0 \\ D[k] & k = k_0 \\ 0 & k < k_0. \end{cases} \quad (2.269)$$

**Proof** Let's compute element  $ij$  of  $H[k, k_0]$ —the response of output coordinate  $i$  to a unit impulse applied at time  $k = k_0$  in input coordinate  $j$ . We'll compute this for each of the three cases above. We first focus on the case of  $k > k_0$ . With  $u[\cdot] = \delta^{k_0}[\cdot]e_j$  the input signal having  $\delta^{k_0}$  in coordinate  $j$  and zeros elsewhere, we have

$$\rho(k, k_0, 0, \delta^{k_0}[\cdot]e_j)_i = \left[ C[k] \sum_{j=k_0}^{k-1} \Phi[k, j+1]B[j]e_j\delta^{k_0}[j] + D[k]e_j\delta^{k_0}[k] \right]_i \quad (2.270)$$

$$= \left[ C[k]\Phi[k, k_0 + 1]B[k_0]e_j \right]_i, \quad (2.271)$$

since for all indices  $j \neq k_0$  in the sum,  $\delta^{k_0}[j]$  equals zero, and for  $j = k_0$ ,  $\delta^{k_0}[j]$  equals one. Now, we recognize this term as,

$$= \left[ C[k] \Phi[k, k_0 + 1] B[k_0] \right]_{ij}, \quad (2.272)$$

since multiplication by  $e_j$  selects column  $j$  of  $C[k] \Phi[k, k_0 + 1] B[k_0]$ . This confirms the formula for  $k > k_0$ . Now, we verify the formula for  $k = k_0$ . We have, for the same  $u[\cdot]$  as above,

$$\rho(k_0, k_0, \delta^{k_0}[\cdot] e_j)_i = \left[ C[k_0] \sum_{j=k_0}^{k_0-1} \Phi[k_0, j+1] B[j] e_j \delta^{k_0}[j] + D[k_0] e_j \delta^{k_0}[k_0] \right]_i \quad (2.273)$$

$$= \left[ D[k_0] e_j \right]_i \quad (2.274)$$

$$= \left[ D[k_0] \right]_{ij}, \quad (2.275)$$

where the sum is empty (and therefore zero) since  $k = k_0$ ,  $u[k_0]$  picks out column  $j$  of  $D[k_0]$ , and  $[\cdot]_i$  picks out row  $i$ . This again matches the given formula. The final component,  $k < k_0 \Rightarrow H[k, k_0] = 0$ , simply follows from the definition of the impulse response.  $\square$

Now that we've established a few basic facts concerning the impulse response, we can return to the goal we outlined above—characterizing the zero-state response of a linear system to *any* signal via a simple, direct input-output relationship. A key step in accomplishing this goal is in recognizing the following fact.

**Lemma 2.12 (Signal Representation via Impulses)** *Consider a signal  $u[\cdot] : \mathbb{Z}_{\geq k_0} \rightarrow \mathbb{R}^m$ . For any  $k \geq k_0$ ,  $u[k]$  can be written as a sum of impulses,*

$$u[k] = \sum_{j=k_0}^k \delta^j[k] u[j] = \sum_{j=k_0}^k \delta[k-j] u[j]. \quad (2.276)$$

*Remark 2.28* The notation  $\mathbb{Z}_{\geq k_0}$  refers to the set of integers that are greater than or equal to a fixed integer  $k_0 \in \mathbb{Z}$ .

**Exercise 2.20** Verify the claim of Lemma 2.12.

Now that we know we can represent *any* discrete-time signal using impulses, we can use impulse response to characterize the zero-state response of a system to any input.

**Theorem 2.13 (Zero-State Response via Impulse Response)** *Consider a discrete-time linear system with I/O map  $\rho$  and impulse response map  $H[\cdot, \cdot] : \mathbb{Z} \times \mathbb{Z} \rightarrow \mathbb{R}^{p \times m}$ . For any input signal  $u : \mathbb{Z}_{\geq k_0} \rightarrow \mathbb{R}^m$ , the zero-state response of the system is computed,*

$$\rho(k, k_0, 0, u[\cdot]) = \sum_{j=k_0}^k H[k, j] u[j]. \quad (2.277)$$

*Remark 2.29* We should be a little careful when defining  $u : \mathbb{Z}_{\geq k_0} \rightarrow \mathbb{R}^m$ ! Formally, we defined the input signal space of a discrete-time system to be  $\mathcal{U} = \{u : \mathbb{Z} \rightarrow \mathbb{R}^m\}$ , so we shouldn't accept signals defined only on  $\mathbb{Z}_{\geq k_0}$ . As such, when we define a signal  $u[\cdot] : \mathbb{Z}_{\geq k_0} \rightarrow \mathbb{R}^m$ , we will identify it with the “extended” input signal,

$$u_e[k] = \begin{cases} u[k] & k \geq k_0 \\ 0 & k < k_0, \end{cases} \quad (2.278)$$



which is defined on all of  $\mathbb{Z}$ . Here, by “identify,” we mean that we will assume  $u$  is equal to  $u_e$  without explicit mention. Since the future behavior of a causal I/O system only depends on the current state and inputs (as a consequence of the restriction axiom), this is a harmless identification to make. Under the identification with  $u_e$ , the input  $u$  formally belongs to  $\mathcal{U}$ , and can be passed into the I/O map  $\rho$  without trouble.

Now, we return to the proof of Theorem 2.13.

**Proof** The proof of this result follows from an application of linearity of the I/O map. Here, to keep the notation simple, we’ll prove the result in the case where  $y \in \mathbb{R}$  and  $u \in \mathbb{R}^m$ —this is sufficient to show the general case, which consists of assembling the full output vector  $y \in \mathbb{R}^p$  by stacking scalar,  $y_i$  terms. With that said, let’s get started on the proof. In the case where  $y \in \mathbb{R}^m$  and  $u \in \mathbb{R}$ , Lemma 2.12 tells us that, for  $u[\cdot] : \mathbb{Z}_{\geq k_0} \rightarrow \mathbb{R}^m$ , we can write,

$$u[k] = \sum_{j=k_0}^k \delta^j[k] u[j] = \sum_{j=k_0}^k \sum_{i=1}^m \delta^j[k] u_i[j] e_i, \quad (2.279)$$

where each  $e_i$  is the  $i$ ’th standard basis vector of  $\mathbb{R}^m$ , with a 1 in index  $i$  and zeros elsewhere. Now, we apply the linearity of the I/O map. We have,

$$\rho(k, k_0, 0, u[\cdot]) = \sum_{j=k_0}^k \sum_{i=1}^m \rho(k, k_0, 0, \delta^j[k] e_i) u_i[j] \quad (2.280)$$

$$= \sum_{j=k_0}^k \sum_{i=1}^m \rho(k, j, 0, \delta^j[k] e_i) u_i[j] \quad (2.281)$$

$$= \sum_{j=k_0}^k \sum_{i=1}^m h_{1i}(k, j) u_i[j] \quad (2.282)$$

$$= \sum_{j=k_0}^k [h_{11}[k, j] \dots h_{1m}[k, j]] u[j] \quad (2.283)$$

$$= \sum_{j=k_0}^k H[k, j] u[j], \quad (2.284)$$

where we make the final equality under the assumption that  $p = 1$ . The general case follows by stacking the formula above,  $p$  times.  $\square$

This tells us that we can *entirely* characterize the zero-state response of a discrete-time linear system to *any* input from knowledge of its impulse response! Further, this impulse response formula is significantly simpler than the complex, I/O map formula we stated above.

Now, let’s specialize what we’ve learned from the time-varying case to the time-invariant case. Due to time-invariance, the impulse response map  $H[k, k_0]$  of an LTI system representation will depend only on the amount of time that has passed,  $k - k_0$ , rather than the objective times  $k$  and  $k_0$ . Thus, we expect the impulse response map of a linear, time-invariant system to satisfy,

$$H[k, k_0] = H[k - k_0, 0] \quad \forall k, k_0 \in \mathbb{Z} \text{ (LTI case)}, \quad (2.285)$$

since the differences between the first and second arguments are identical in the left and right-hand sides. Consequently, in order to characterize the impulse response of an LTI system, it's sufficient to know  $H[k, 0]$  for all  $k \in \mathbb{Z}$ , since  $H[k, k_0]$  will equal  $H[k - k_0, 0]$ . We introduce the following special “definition,” (which is really just special notation) to highlight this fact.

**Definition 2.31 (DT-LTI Impulse Response Map)** Consider a discrete-time, LTI system with impulse response map  $H[\cdot, \cdot] : \mathbb{Z} \times \mathbb{Z} \rightarrow \mathbb{R}^{p \times m}$ . The LTI impulse response map of the system is the map  $H[\cdot] : \mathbb{Z} \rightarrow \mathbb{R}^{p \times m}$ , defined

$$H[k] := H[k, 0], \quad \forall k \in \mathbb{Z}. \quad (2.286)$$

*Remark 2.30* In the definition above, we use the same letter for  $H[\cdot, \cdot]$  and  $H[\cdot]$ . Although this *is* admittedly an overload of notation, one should always be able to determine from context which impulse response map is being used. If only one argument appears, assume the LTI impulse response map is being used, while if two arguments appear, assume the general (LTV) impulse response map is being used. Just like we use the same letter to refer to both maps, we'll also freely refer to both the (LTV) impulse response map and the LTI impulse response map as *the* impulse response map of the system. Again, context should make clear which map we're referring to.

Now, we compute the impulse response map of a discrete-time, LTI system.

**Proposition 2.22 (Impulse Response Map of a DT-LTI System Representation)**

*Consider a discrete-time, LTI system representation  $(A, B, C, D)$ . The LTI impulse response map  $H[\cdot] : \mathbb{Z} \rightarrow \mathbb{R}^{p \times m}$  of the system is computed,*

$$H[k] = \begin{cases} CA^{k-1}B & k > 0 \\ D & k = 0 \\ 0 & k < 0. \end{cases} \quad (2.287)$$

**Exercise 2.21** Using the formula for the discrete-time, LTI state transition matrix, verify the formula proposed in Proposition 2.22.

Using the LTI impulse response map  $H[\cdot] : \mathbb{Z} \rightarrow \mathbb{R}^{p \times m}$  in conjunction with Theorem 2.13, we observe that we can write the zero-state response  $y[k] = \rho(k, k_0, 0, u[\cdot])$  of a discrete-time LTI system to an input  $u : \mathbb{Z}_{\geq k_0} \rightarrow \mathbb{R}^m$  as the sum,

$$y[k] = \sum_{j=k_0}^k H[k, j]u[j] = \sum_{j=k_0}^k H[k-j]u[j] = \sum_{j=-\infty}^{\infty} H[k-j]u[j], \quad (2.288)$$

where we extend the lower bound to  $-\infty$  using our convention that  $u[k] = 0$  for  $k < k_0$  and extend the upper bound to  $+\infty$  using  $H[k-j] = 0$  for  $k-j < 0$ . Sums of this form are of *fundamental* importance in control theory, signal processing, and beyond.

**Definition 2.32 (Convolution Sum)** Consider two discrete-time signals,  $u[\cdot]$  and  $H[\cdot]$ , of compatible dimension, for which  $H[k] \cdot u[k] \in \mathbb{R}^p$ . The *convolution* of  $H$  and  $u$  is the signal  $(H * u)[\cdot] : \mathbb{Z} \rightarrow \mathbb{R}^p$ , defined as the sum,

$$(H * u)[k] = \sum_{j=-\infty}^{\infty} H[k-j]u[j], \quad \forall k \in \mathbb{Z}. \quad (2.289)$$

Such a sum is called a *convolution sum*.

*Remark 2.31* It's critical to note - the convolution of two signals is *another signal*, not a fixed value! That is,  $H * u$  is a map from  $\mathbb{Z} \rightarrow \mathbb{R}^p$ , not a single vector in  $\mathbb{R}^p$ .

*Remark 2.32* Since the convolution sum is an *infinite sum*, special care should be taken to ensure that the sum converges before using it.

*Remark 2.33* If one takes the convolution of signals  $H$  and  $u$ , one says that  $H$  and  $u$  are being *convolved* (not *convoluted*).

*Remark 2.34* We briefly elaborate on what we mean by “compatible dimension.” Frequently, we'll convolve a matrix-valued signal  $H[\cdot] : \mathbb{Z} \rightarrow \mathbb{R}^{p \times m}$  with a vector-valued signal,  $u[\cdot] : \mathbb{R} \rightarrow \mathbb{R}^m$ . Since the product  $H[k] \cdot u[k]$  is well-defined for  $H[k] \in \mathbb{R}^{p \times m}$  and  $u[k] \in \mathbb{R}^m$ , we say that the signals are of compatible dimension. Sometimes, we'd also like to convolve a scalar signal  $f[\cdot] : \mathbb{Z} \rightarrow \mathbb{R}$  with a vector signal  $u[\cdot] : \mathbb{Z} \rightarrow \mathbb{R}^m$ . Despite  $f[k]$  not having  $m$  rows, the product  $f[k] \cdot u[k]$  is still well-defined, so we still say that the dimensions are compatible. It's for this reason that we don't require  $H[k] \in \mathbb{R}^{p \times m}$  and  $u[k] \in \mathbb{R}^m$ —asking for *compatible dimension* yields greater flexibility.

By applying Lemma 2.12 and the definition of a convolution sum, we can rephrase the signal representation property of the discrete-time impulse in the language of convolution.

**Lemma 2.13 (Impulse is the Identity of Convolution)** *For any signal  $u[\cdot] : \mathbb{Z} \rightarrow \mathbb{R}^p$  and the discrete-time impulse  $\delta[\cdot] : \mathbb{Z} \rightarrow \mathbb{R}^p$ , it follows that  $u[\cdot] = (\delta * u)[\cdot]$ .*

**Proof** Immediate from Lemma 2.12 and Definition 2.32.  $\square$

Applying this lemma together with Theorem 2.13, we arrive at the following result.

**Corollary 2.2 (Convolution of Input & Impulse Response Equals Output)** *Consider a discrete-time, LTI system with impulse response map  $H[\cdot] : \mathbb{Z} \rightarrow \mathbb{R}^{p \times m}$ . For any input signal  $u[\cdot] : \mathbb{Z}_{\geq 0} \rightarrow \mathbb{R}^m$ , the zero-state response of the system is equal to the convolution of the impulse response and input,*

$$\rho(k, 0, 0, u[\cdot]) = (H * u)[k]. \quad (2.290)$$

**Proof** This result follows from a straightforward application of Theorem 2.13 and Definition 2.32. Since for  $k < 0$ , one has  $H[k] = 0$  (by definition of impulse response), it follows that

$$(H * u)[k] = \sum_{j=-\infty}^{\infty} H[k-j]u[j] = \sum_{j=-\infty}^k H[k-j]u[j] = \sum_{j=0}^k H[k-j]u[j], \quad (2.291)$$

where we use that  $H[k-j]$  is zero for  $k-j < 0$  and that  $u[j]$  is zero for  $j < 0$  (under the convention we established for signals on  $\mathbb{Z}_{\geq 0}$ ). Since  $H[k-j] = H[k, j]$ , we observe that the final expression matches the formula for  $\rho(k, 0, 0, u[\cdot])$  from Theorem 2.13.  $\square$

*Remark 2.35* Recall that, in Section 1 of this chapter, we defined all of our systems to be *causal*. Due to causality,  $H[k - j] = 0$  for  $k - j < 0$ , since causal systems don't require information about the future (time  $j$ , for  $j > k$ ) to determine information about the present (time  $k$ ). As we observed in the proof above, this property ensured the convolution sum did not truly extend to  $+\infty$ . For *noncausal* systems, however, this is not necessarily the case—one must be more careful to ensure that the sum converges.

**Exercise 2.22** So far, we've focused on systems with zero initial condition. Confirm that, for a discrete-time, LTI system representation  $(A, B, C, D)$ , the output of the system from initial state  $x[0] = x_0$  and input signal  $u[\cdot] : \mathbb{Z}_{\geq 0} \rightarrow \mathbb{R}^m$  is computed  $y[k] = C\Phi[k, 0]x_0 + (H * u)[k]$ .

## 2.4.2 Impulse Response of Continuous-Time Systems

*Note: Since a rigorous construction of the continuous-time impulse response involves a digression into distribution theory, we'll be a little bit more hand-wavy than usual in this subsection. We direct the reader to the references at the end of the section for a more detailed mathematical approach to the content covered here.*

Now that we've developed a theory for the discrete-time impulse response, we aim to construct an analogous theory for continuous-time systems. Following the same path we took in the discrete-time case, our first step is to answer the question: what is the correct notion of impulse response for a continuous-time system?

In order to answer this question, we must first define a continuous-time impulse signal. To motivate the definition of this signal, let's recall the important properties of the discrete-time impulse. Perhaps the most important property of the discrete-time impulse  $\delta[\cdot]$  was that any signal  $u[\cdot] : \mathbb{Z} \rightarrow \mathbb{R}^m$  could be written as a *convolution* with the impulse,

$$u[k] = (\delta * u)[k] = \sum_{j=-\infty}^{\infty} \delta[k - j]u[j]. \quad (2.292)$$

If we seek a continuous-time analogue of the discrete-time unit impulse, perhaps we should define a continuous-time analogue of a convolution sum and a continuous-time signal  $\delta$  for which  $u = \delta * u$ .

First, let's define an appropriate notion of continuous-time convolution. Thus far, when translating between discrete and continuous-time, we've swapped sums for integrals wherever they appear. In order to develop a continuous-time convolution, it's therefore natural to replace a convolution sum with a convolution integral.

**Definition 2.33 (Convolution Integral)** Consider two continuous-time signals  $u(\cdot)$  and  $H(\cdot)$  of compatible dimension, for which  $H(t) \cdot u(t) \in \mathbb{R}^p$ . The convolution of  $H$  and  $u$  is the signal  $(H * u)(\cdot) : \mathbb{R} \rightarrow \mathbb{R}^p$ , defined

$$(H * u)(t) = \int_{\mathbb{R}} H(t - \tau)u(\tau)d\tau. \quad (2.293)$$

*Remark 2.36* Above, we make use of the notation  $\int_{\mathbb{R}}$  to indicate that the integral is being taken from  $-\infty$  to  $+\infty$ . As in the case of the convolution sum, one must take special care to

ensure the convolution integral converges before performing any operations, as the integral is being taken over all of  $\mathbb{R}$ .

*Remark 2.37* By “compatible dimension,” we mean the same thing as in the discrete-time case—the dimensions of  $H$  and  $u$  must be such that the product  $H(t) \cdot u(t)$  is defined.

Now that we’ve defined a continuous-time notion of convolution, we can define a continuous-time unit impulse function,  $\delta(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$ . If we translate directly from discrete-time to continuous-time, we find that the continuous-time unit impulse function should satisfy,

$$u(t) = (\delta * u)(t) = \int_{\mathbb{R}} \delta(t - \tau)u(\tau)d\tau, \quad \forall t \in \mathbb{R}, \quad (2.294)$$

for all piecewise continuous signals  $u : \mathbb{R} \rightarrow \mathbb{R}^m$ . What *is* this continuous-time function  $\delta$ ? The following proposition gives quite a worrying answer to this question:  $\delta$  *cannot* exist.

**Proposition 2.23 (Nonexistence of the  $\delta$  Function)** *There is no integrable<sup>7</sup> function  $\delta : \mathbb{R} \rightarrow \mathbb{R}$  such that for all  $u \in PC(\mathbb{R}, \mathbb{R})$ ,  $\int_{\mathbb{R}} \delta(t - \tau)u(\tau)d\tau = u(t)$ ,  $\forall t \in \mathbb{R}$ .*

*Remark 2.38* Notice that we state this proposition for a signal  $u : \mathbb{R} \rightarrow \mathbb{R}$ , rather than a signal  $u : \mathbb{R} \rightarrow \mathbb{R}^m$ . Since integrals of vector-valued functions are evaluated element-wise, the scalar case is sufficient to conclude the vector case.

**Proof (Sketch)** Suppose for contradiction that there exists a function  $\delta : \mathbb{R} \rightarrow \mathbb{R}$  satisfying  $u(t) = \int_{\mathbb{R}} \delta(t - \tau)u(\tau)d\tau$ , for all  $u \in PC(\mathbb{R}, \mathbb{R})$ . By necessity, such a function must be zero almost everywhere<sup>8</sup>—if this were not the case, one would not have  $u(t) = \int_{\mathbb{R}} \delta(t - \tau)u(\tau)d\tau$  for all  $u$  (draw some pictures to convince yourself of this). Now, consider the constant signal,  $u(t) \equiv 1$ . By assumption, for all  $t \in \mathbb{R}$ ,

$$u(t) = 1 = \int_{\mathbb{R}} \delta(t - \tau)u(\tau)d\tau = \int_{\mathbb{R}} \delta(t - \tau)d\tau, \quad (2.295)$$

which implies  $\int_{\mathbb{R}} \delta(\tau)d\tau = 1$ . But, since  $\delta$  is zero almost everywhere, it must be that  $\int_{\mathbb{R}} \delta(\tau)d\tau = 0$ —contradiction! We conclude a  $\delta$  function cannot exist.  $\square$

This presents us with quite the problem! In order to define the impulse response of a continuous-time system, we need a continuous-time impulse signal with the correct convolution property. Yet, we’ve just shown that no such signal exists!

Consequently, in order to provide a proper definition of the continuous-time impulse, we must look beyond simple signals and turn to the theory of *distributions*, or “generalized signals,” as they’re sometimes called. We now provide an informal definition of a special distribution, called the *Dirac delta distribution*, which has the property we’re looking for.

**Definition 2.34 ((Informal) Dirac Delta Distribution)** Let  $\mathcal{E} \subseteq \{f : \mathbb{R} \rightarrow \mathbb{C}\}$  be a vector space of complex-valued functions, called *test functions*. The Dirac delta distribution on  $\mathcal{E}$  is a linear map  $\delta : \mathcal{E} \rightarrow \mathbb{C}$ , equipped with an operation  $*$ , called convolution, for which  $\delta$  and  $*$  satisfy the following:

<sup>7</sup> Here, by *integrable*, we mean a function whose magnitude has a well-defined Lebesgue integral over  $\mathbb{R}$ . If you’re unfamiliar with the Lebesgue integral, feel free to treat it as a fancy-sounding Riemann integral—we won’t focus on the technical details of integration theory in this course.

<sup>8</sup> A function is zero *almost everywhere* if it is zero everywhere but on a set of Lebesgue measure zero. Lebesgue measure zero sets are small sets (e.g.  $\{0\}$ ,  $\mathbb{N}$ , etc.) that are “ignored” by Lebesgue integrals.

1. Origin:  $\delta(f) = f(0)$  for all  $f \in \mathcal{E}$ .
2. Impulse:  $(\delta * f)(t) = f(t)$ , for all  $t \in \mathbb{R}$  and  $f \in \mathcal{E}$ .
3. Linearity:  $(\delta * (\alpha f + \beta g))(t) = \alpha f(t) + \beta g(t)$ , for all  $t \in \mathbb{R}$ ,  $\alpha, \beta \in \mathbb{C}$ , and  $f, g \in \mathcal{E}$ .
4. Associativity:  $\delta * (f * g) = (\delta * f) * g$  for all  $f, g \in \mathcal{E}$ , where  $f * g$  represents the (standard) integral convolution of  $f$  and  $g$ .

We define the *shifted* Dirac delta distribution to be the map  $\delta^{t_0} : \mathcal{E} \rightarrow \mathbb{C}$ , satisfying  $\delta^{t_0}(f) = f(t_0)$  and  $(\delta^{t_0} * f)(t) = f(t - t_0)$ , where  $t_0 \in \mathbb{R}$  is a fixed time.

*Remark 2.39* It's *critical* to note—the Dirac delta distribution is *not* a signal! Instead of taking in *time* (like a signal might), it directly takes in *signals* (from the space of test functions). It is this aspect of the definition that lets us work around the problem posed by Proposition 2.23. We have simply defined the Dirac delta distribution to be a map on a space of signals that gives us exactly the properties we want.

*Remark 2.40* There are a few things that require sharpening in order to make this definition formal. First, the space  $\mathcal{E}$  of test functions is usually chosen to have some “nice” structure. For instance,  $\mathcal{E}$  is often taken to be a set of  $C^\infty$  functions with compact support. Secondly, convolution should be defined for *all* distributions, not just for the Dirac  $\delta$ , and should be confirmed to interact well with the standard integral convolution. There are a couple of other technical points—regarding convergence and continuity properties—that must be satisfied to form a proper definition. We direct the interested reader to the references at the end of the chapter for these details.

Properly motivating and studying distribution theory takes quite a bit of work—as such, in the remainder of this section, we'll give a semi-formal treatment of the continuous-time impulse response using the informal definition of the Dirac delta that we posed above.

In many engineering texts, one will find the Dirac delta written as a *function of time*,  $\delta(t)$ . Although this is not correct in the formal mathematical sense, in situations where rigor is *not* a concern, manipulating  $\delta$  as if it were a function of time can help in gaining some basic insight into problems. As such, we quickly lay out a few *informal* ground rules for working with a Dirac delta as if it were a function of time. We stress—these rules *should not* be used in a rigorous mathematical context—the Dirac delta distribution is a map on a space of functions, *not* a regular function of time! In the remainder of this section, we'll be precise in pointing out where we use these “function of time” formulas.

**Proposition 2.24 ((Informal) Dirac Delta as a Function of Time)** *Let  $u \in PC(\mathbb{R}, \mathbb{R})$ . The following informal “rules” are used to manipulate the Dirac delta as a function of time. For all  $t$  (or  $t - t_0$  for the second formula) at which  $u(\cdot)$  is continuous,*

$$u(t) = (\delta * u)(t) \text{ “=” } \int_{\mathbb{R}} \delta(t - \tau)u(\tau)d\tau \quad (2.296)$$

$$u(t - t_0) = (\delta^{t_0} * u)(t) \text{ “=” } \int_{\mathbb{R}} \delta^{t_0}(t - \tau)u(\tau)d\tau. \quad (2.297)$$

*If  $u(\cdot)$  is continuous at 0 (or at  $t_0$  for the second formula), then*

$$u(0) = \delta(u) \text{ “=” } \int_{\mathbb{R}} \delta(\tau)u(\tau)d\tau \quad (2.298)$$

$$u(t_0) = \delta^{t_0}(u) \text{ “=” } \int_{\mathbb{R}} \delta^{t_0}(\tau)u(\tau)d\tau. \quad (2.299)$$

*Remark 2.41* It's essential to remember: the integral “rules,” written with equality in quotations, are *not* mathematically rigorous! It's best to think of each “rule” as *notation*, rather than as mathematical fact. In the optional, starred subsection following this, we will use a technique called *approximations to the identity* to more rigorously justify the use of these informal integral formulas.

*Remark 2.42* Here, we've asked that each property hold at a time where  $u(\cdot)$  is continuous. This request is a consequence of how integrals deal with discontinuity. If  $u$  had a jump discontinuity at a single point, the integral would simply *ignore* the jump discontinuity, since integrals are agnostic to changes on single-point sets. Thus, if we ask that  $u(t) = \int_{\mathbb{R}} \delta(t - \tau)u(\tau)$ , we should at best expect the equality to hold at points of continuity, since the integral “filters out” single point discontinuities. The idea of integrals “filtering out” function values on certain, small sets is the starting point of a rigorous theory of integration via measure theory.

**Proof (Informal)** These “rules” are motivated as follows. The first two formulas follow from the definition of convolution with the Dirac delta,

$$u(t) = (\delta * u)(t), \quad u(t - t_0) = (\delta^{t_0} * u)(t). \quad (2.300)$$

Here, we just swap out the abstract convolution operation we defined for distributions for the integral definition of convolution. The second two formulas follow from the first two convolution integrals, and are analogous to the properties  $u(0) = \delta(u)$  and  $u(t_0) = \delta^{t_0}(u)$ , which we stated in the definition above.  $\square$

Now that we've sketched out a basic, informal definition for the continuous-time impulse “signal,” we can return to the problem of defining impulse response for a continuous-time system. Recall that, for a discrete-time linear, time-varying system, the impulse response map was defined as a map  $H : \mathbb{Z} \times \mathbb{Z} \rightarrow \mathbb{R}^{p \times m}$ , in which  $[H[k, k_0]]_{ij}$  was the zero-state response of output coordinate  $i$  to an impulse applied in input coordinate  $j$  at time  $k_0$ .

Using the discrete-time definition as a guide, let's motivate a definition for continuous-time impulse response. Recall that the zero-state response of a continuous-time linear, time-varying system is computed,

$$\rho(t, t_0, 0, u(\cdot)) = C(t) \int_{t_0}^t \Phi(t, \tau) B(\tau) u(\tau) d\tau + D(t) u(t). \quad (2.301)$$

Let's calculate the output of the system for the input  $\delta^{t_0} e_j$ , a time-shifted Dirac delta applied in input coordinate  $j$ . Using the informal integral rules we established above, we have,

$$\rho(t, t_0, 0, \delta^{t_0} e_j) = C(t) \int_{t_0}^t \Phi(t, \tau) B(\tau) e_j \delta^{t_0}(\tau) d\tau + D(t) e_j \delta^{t_0} \quad (2.302)$$

$$= C(t) \Phi(t, t_0) B(t_0) e_j + D(t) e_j \delta^{t_0}. \quad (2.303)$$

Taking the  $i$ 'th coordinate, we find that the zero-state response of coordinate  $i$  is,

$$\rho(t, t_0, 0, \delta^{t_0} e_j)_i = [C(t) \Phi(t, t_0) B(t_0) + D(t) \delta^{t_0}]_{ij}. \quad (2.304)$$

This motivates the following definition for the continuous-time linear, time-varying impulse response map.

**Definition 2.35 (CT-LTV Impulse Response Map)** Consider a continuous-time linear, time-varying system representation  $(A(\cdot), B(\cdot), C(\cdot), D(\cdot))$ . The impulse response map  $H(\cdot, \cdot)$  of the system representation is defined on  $\mathbb{R} \times \mathbb{R}$  as,

$$H(t, t_0) = \begin{cases} C(t)\Phi(t, t_0)B(t_0) + D(t)\delta^{t_0} & t \geq t_0 \\ 0 & t < t_0. \end{cases} \quad (2.305)$$

Entry  $[H(t, t_0)]_{ij} = h_{ij}(t, t_0)$  is interpreted as the  $i$ 'th component of the zero-state response to an input  $\delta^{t_0}e_j$  containing a shifted Dirac delta at  $t_0$  in its  $j$ 'th component.

*Remark 2.43* Due to the presence of the Dirac delta when  $D(\cdot)$  is nonzero,  $H$  is *not* a map that can be evaluated like a “normal” function into  $\mathbb{R}^{p \times m}$ , since  $\delta^{t_0}$  is not something we can evaluate as a function of time. Rather,  $H$  is a tool that we will use in conjunction with integration and convolution—operations in which we understand how  $\delta^{t_0}$  behaves.

Now, let's show that we can use the impulse response map to compute the zero-state response of the system to any (admissible) input. What should we expect the formula for zero-state response to look like? Recall that, for a discrete-time linear, time-varying system, the zero-state response of the system to an input  $u[\cdot]$  was computed  $\rho(k, k_0, 0, u[\cdot]) = \sum_{j=k_0}^k H[k, j]u[j]$ . Translating this *sum* formula into an *integral* formula, we expect the continuous-time case to satisfy,

$$\rho(t, t_0, 0, u(\cdot)) = \int_{t_0}^t H(t, \tau)u(\tau)d\tau. \quad (2.306)$$

The following theorem confirms that this formula holds at every  $t$  at which the problem data is continuous.

**Theorem 2.14 (Impulse Response of a Linear, Time-Varying System)** Consider an LTV system representation  $(A(\cdot), B(\cdot), C(\cdot), D(\cdot))$  with impulse response map  $H(\cdot, \cdot)$ . For any  $u \in PC(\mathbb{R}_{\geq t_0}, \mathbb{R}^m)$  and  $t \in \mathbb{R}$  at which  $A(\cdot), B(\cdot), C(\cdot), D(\cdot), u(\cdot)$  are continuous,

$$\rho(t, t_0, 0, u(\cdot)) = \int_{t_0}^t H(t, \tau)u(\tau)d\tau. \quad (2.307)$$

*Remark 2.44* Just like in the discrete-time case, when we define a signal  $u \in PC(\mathbb{R}_{\geq t_0}, \mathbb{R}^m)$ , we will automatically identify it with the extended signal  $u_e \in PC(\mathbb{R}, \mathbb{R}^m)$ , which is zero for  $t < t_0$  and equal to  $u$  for  $t \geq t_0$ . This way, an input signal  $u \in PC(\mathbb{R}_{\geq t_0}, \mathbb{R}^m)$  can be passed into the I/O map  $\rho$  without trouble.

**Proof (Informal)** We'll provide an informal proof of this result using the informal “rules” for manipulating  $\delta$  as if it were a function of time. In particular, we will use the “rule” that, for every  $t$  at which  $u(\cdot)$  is continuous,

$$u(t) = \int_{\mathbb{R}} \delta(t - \tau)u(\tau)d\tau. \quad (2.308)$$

Since we are given  $u : \mathbb{R}_{\geq t_0} \rightarrow \mathbb{R}^m$ , we will perform our standard identification of  $u$  with a signal  $u_e : \mathbb{R} \rightarrow \mathbb{R}^m$ , defined on all of  $\mathbb{R}$ , by setting  $u(t) = 0$  for  $t < t_0$ . Plugging into the formula for the zero-state response of a CT-LTV system yields,



$$\rho(t, t_0, 0, u(\cdot)) = C(t) \int_{t_0}^t \Phi(t, \tau) B(\tau) u(\tau) d\tau + D(t) u(t) \quad (2.309)$$

$$= C(t) \int_{-\infty}^t \Phi(t, \tau) B(\tau) u(\tau) d\tau + D(t) u(t), \quad (2.310)$$

since the input signal is taken to be zero for  $t < t_0$ . Letting  $\mathbb{1}_{\tau \leq t}(\tau)$  denote the indicator function which is equal to 1 when  $\tau \leq t$  and zero elsewhere, it follows that,

$$= \int_{\mathbb{R}} \mathbb{1}_{\tau \leq t}(\tau) C(t) \Phi(t, \tau) B(\tau) u(\tau) d\tau + D(t) \int_{\mathbb{R}} \delta(t - \tau) u(\tau) d\tau \quad (2.311)$$

$$= \int_{\mathbb{R}} \mathbb{1}_{\tau \leq t}(\tau) C(t) \Phi(t, \tau) B(\tau) u(\tau) d\tau + \int_{\mathbb{R}} \mathbb{1}_{\tau \leq t}(\tau) D(t) \delta(t - \tau) u(\tau) d\tau \quad (2.312)$$

$$= \int_{\mathbb{R}} \mathbb{1}_{\tau \leq t}(\tau) [C(t) \Phi(t, \tau) B(\tau) + D(t) \delta(t - \tau)] u(\tau) d\tau, \quad (2.313)$$

where we use our informal integral rules to add the indicator to the  $D(t)$  integral. Now, we recognize the term in brackets as  $H(t, \tau)$ . We therefore have,

$$\rho(t, t_0, 0, u(\cdot)) = \int_{\mathbb{R}} \mathbb{1}_{\tau \leq t}(\tau) H(t, \tau) u(\tau) d\tau \quad (2.314)$$

$$= \int_{\tau \leq t} H(t, \tau) u(\tau) d\tau \quad (2.315)$$

$$= \int_{t_0}^t H(t, \tau) u(\tau) d\tau, \quad (2.316)$$

where in the last step, we use that  $u(\tau)$  is zero for  $\tau \leq t_0$ .  $\square$

Next, we specialize to the LTI case. As with the discrete-time case, we recognize that time passed, rather than objective start and end times, is the relevant quantity.

**Definition 2.36 (CT-LTI Impulse Response Map)** Consider a continuous-time, LTI system with impulse response map  $H(\cdot, \cdot)$  on  $\mathbb{R} \times \mathbb{R}$ . The LTI impulse response map of the system is the map  $H(\cdot)$  on  $\mathbb{R}$ , defined

$$H(t) = H(t, 0), \quad \forall t \in \mathbb{R}. \quad (2.317)$$

As with the discrete-time case, this map satisfies  $H(t, t_0) = H(t - t_0)$  due to time-invariance. We now compute the impulse response map of a continuous-time, LTI system, and prove that the zero-state response of such a system can be computed via a convolution.

**Corollary 2.3 (Impulse Response of a Linear, Time-Varying System)** Consider a continuous-time, LTI system representation  $(A, B, C, D)$ . The LTI impulse response map of the system is computed,

$$H(t) = \begin{cases} C \exp(At) B + D \delta & t \geq 0 \\ 0 & t < 0. \end{cases} \quad (2.318)$$

For any  $u \in PC(\mathbb{R}_{\geq 0}, \mathbb{R}^m)$  and any  $t \in \mathbb{R}$  at which  $u(\cdot)$  is continuous,

$$\rho(t, 0, 0, u(\cdot)) = (H * u)(t). \quad (2.319)$$

**Proof** The formula for  $H$  follows directly from Definition 2.35 and application of the formula  $\Phi(t, t_0) = \exp(A(t - t_0))$  for a continuous-time, LTI system. The convolution formula for the output response follows directly from Theorem 2.14 and the definition of continuous-time convolution. We have, for  $u(\cdot) \in PC(\mathbb{R}_{\geq 0}, \mathbb{R}^m)$  and any  $t \geq 0$  at which  $u(\cdot)$  is continuous,

$$\rho(t, 0, 0, u(\cdot)) = \int_0^t H(t - \tau)u(\tau)d\tau = \int_{\mathbb{R}} H(t - \tau)u(\tau)d\tau = (H * u)(t). \quad (2.320)$$

Thus, the desired formula holds.  $\square$

### 2.4.3 Approximations to the Identity ★

Above, we gave an informal introduction to the continuous-time impulse response using an informal definition of the Dirac delta. In doing this, we introduced a number of “rules” for manipulating the Dirac delta as if it were a function of time. In this optional subsection, we provide a firmer theoretical justification for these “rules,” and see that we can *approximate* the behavior of the Dirac delta distribution with a sequence of well-behaved functions. Note that this section can be skipped without much loss of continuity—however, skimming the basic ideas and looking at the figures could be helpful for your understanding. Let’s get started!

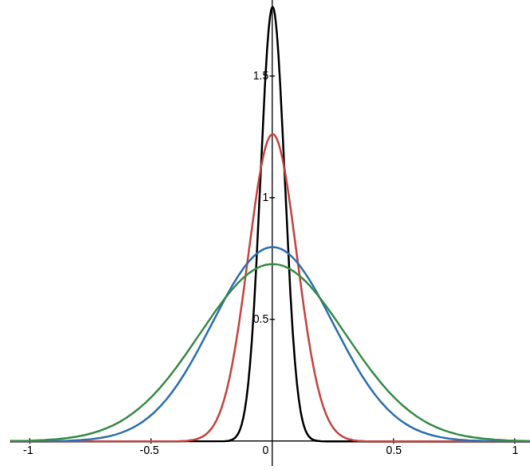
Above, we proved that there is *no* continuous-time signal which has the convolution property of the Dirac delta. In order to get around this problem, one may take two approaches. First, one may take the approach we hinted at above, and learn some (formidable) distribution theory. Second, instead of studying the Dirac delta via distribution theory, one can try to *approximate* the Dirac delta using a sequence of continuous-time signals that—in the limit—give us the same convolution behavior as the *distribution*  $\delta$ . Here, we’ll provide a brief overview of the second approach. Notably, by approximating  $\delta$  with a family of well-behaved signals, we’ll be able to draw connections between the informal integration “rules” we outlined above and formal mathematical facts.

How do we approximate the Dirac delta *distribution* (which is a map on a space of functions) with continuous-time signals? Let’s turn to the informal integral “rules” for manipulating the Dirac delta as if it were a signal to get some ideas. Ideally, we want a family of approximations,  $\delta_\epsilon : \mathbb{R} \rightarrow \mathbb{R}$ , parameterized by  $\epsilon > 0$ , such that for every admissible signal  $u(\cdot)$  and sufficiently small  $\epsilon$ ,

$$(\delta_\epsilon * u)(t) = \int_{\mathbb{R}} \delta_\epsilon(t - \tau)u(\tau)d\tau \approx u(t). \quad (2.321)$$

Further, we’d like this approximation to become exact as  $\epsilon \rightarrow 0$ . In order for this to happen, it seems like  $\delta_\epsilon(t - \tau)$  should have a high value near  $t$ , and small values elsewhere. That is,  $\delta_\epsilon(t - \tau)$  should have a *sharp peak* near  $t$  and should drop off to zero elsewhere—this way, the integral will “pick out” the value of  $u(\cdot)$  at time  $t$ . With this in mind, we define a class of “well-behaved” approximations of  $\delta$ , called *approximations to the identity*.

**Definition 2.37 (Approximation to the Identity)** An approximation to the identity is a collection  $\{\delta_\epsilon\}_{\epsilon > 0}$  of integrable functions  $\delta_\epsilon : \mathbb{R} \rightarrow \mathbb{R}$ , for which the following are satisfied:



**Fig. 2.5** A series of approximations of the  $\delta$  “function,” formed by Gaussian distributions of successively smaller variances, which place more and more mass at the origin as  $\epsilon \rightarrow 0$ . Once  $\epsilon$  “equals” zero, all of the mass is placed at the origin, and the approximation appears to be a “spike” that jumps to  $\infty$  at the origin.

1. Unit mass: for all  $\epsilon > 0$ ,  $\int_{\mathbb{R}} \delta_{\epsilon}(t) dt = 1$ .
2. Uniform bound: there exists an  $A > 0$  such that for all  $\epsilon > 0$ ,  $\int_{\mathbb{R}} |\delta_{\epsilon}(t)| dt \leq A$ .
3. Limiting behavior: for every  $\eta > 0$ ,  $\int_{|t| \geq \eta} |\delta_{\epsilon}(t)| dt \rightarrow 0$  as  $\epsilon \rightarrow 0$ .

For  $t_0 \in \mathbb{R}$ , an approximation to the identity at  $t_0$  is a family  $\{\delta_{\epsilon}^{t_0}\}_{\epsilon > 0}$  for which  $\{\delta_{\epsilon}\}_{\epsilon > 0}$ ,  $\delta_{\epsilon}(t) := \delta_{\epsilon}^{t_0}(t + t_0)$ , is an approximation to the identity.

*Remark 2.45* As we’ll find out below, this definition is sufficient to enable pointwise convergence theorems for convolutions with piecewise continuous functions. If one wants stronger convergence modes or convergence results on more general classes of signals, stronger versions of (2) and (3) should be imposed—see Chapter 3.2 of [37] for further details.<sup>9</sup>

*Remark 2.46* The name *approximation to the identity* follows from the fact that  $\delta$  is the “identity” of the convolution,  $\delta * u = u$ . Since approximations to the identity yield approximations of  $\delta$ , it follows that they approximate the “identity” of the convolution operation.

As demonstrated in Figure 2.5, approximations to the identity shift all of their mass to the origin as  $\epsilon \rightarrow 0$ . Since approximations to the identity must *also* satisfy  $\int_{\mathbb{R}} \delta_{\epsilon}(t) dt = 1$ , this shifting property implies that an approximation must “spike” towards infinity at 0 as  $\epsilon \rightarrow 0$ . This is where the *heuristic* definition of a continuous-time impulse function as the spike,

$$\delta(t) \text{ “=” } \begin{cases} \infty & t = 0 \\ 0 & t \neq 0, \end{cases} \quad (2.322)$$

originates from. Although is a nice picture to have in mind when reasoning about the continuous-time impulse at a non-rigorous level, one should always remember—this is *not*

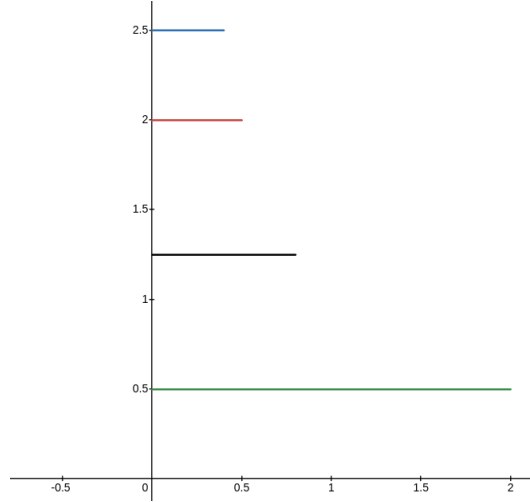
<sup>9</sup> The definition above is referred to as a *good kernel* in [37] and as an *approximation to the identity* in [39]. Here, our definition is consistent with [39].

the definition of  $\delta$ —Proposition 2.23 tells us that a  $\delta$  signal cannot exist. The Dirac delta is truly a map on a function space, *not* a continuous-time signal. Let's consider a simple example of an approximation to the identity.

*Example 2.3* Consider the family of functions,  $\delta_\epsilon : \mathbb{R} \rightarrow \mathbb{R}$ , defined

$$\delta_\epsilon(t) = \begin{cases} 0 & t < 0 \\ 1/\epsilon & t \in [0, \epsilon] \\ 0 & t > \epsilon. \end{cases} \quad (2.323)$$

We now verify this satisfies the properties of an approximation to the identity. We have that  $\int_{\mathbb{R}} \delta_\epsilon(t) dt = 1$  for all  $\epsilon > 0$  and that  $\int_{\mathbb{R}} |\delta_\epsilon(t)| dt = 1$ , from which we conclude the first two conditions. Now, fix  $\eta > 0$ . For  $\eta > \epsilon$ ,  $\int_{|t| \geq \eta} |\delta_\epsilon(t)| dt = 0$ , which implies that for every  $\eta > 0$ ,  $\int_{|t| \geq \eta} |\delta_\epsilon(t)| dt \rightarrow 0$  as  $\epsilon \rightarrow 0$ . Thus, the family  $\{\delta_\epsilon\}_{\epsilon > 0}$  is an approximation to the identity.



**Fig. 2.6** A second approximation to the identity, formed from  $\delta_\epsilon(t) = 1/\epsilon$ , for  $t \in [0, \epsilon]$ .

We now apply these approximations to the identity to study the convolution integral.

**Theorem 2.15 (Approximations to the Identity Converge to  $\delta$ )** *Let  $u \in PC(\mathbb{R}, \mathbb{R}^m)$ . For  $\{\delta_\epsilon\}_{\epsilon > 0}$  an approximation to the identity, at every  $t \in \mathbb{R}$  at which  $u$  is continuous,*

$$\lim_{\epsilon \rightarrow 0} \int_{\mathbb{R}} \delta_\epsilon(t - \tau) u(\tau) d\tau = \lim_{\epsilon \rightarrow 0} (\delta_\epsilon * u)(t) = u(t). \quad (2.324)$$

*Remark 2.47* This property rigorously justifies our integral “rules” for treating the Dirac delta function as a continuous-time signal. Recall that earlier, we stated that if one manipulates  $\delta$  as if it were a function of time, one has the “rule,”

$$\int_{\mathbb{R}} \delta(t - \tau) u(\tau) d\tau = u(t), \quad (2.325)$$

at every  $t$  at which  $u(\cdot)$  is continuous. Now, we observe that this “rule” is really just special notation for the limiting case of Theorem 2.15. Theorem 2.15 therefore provides theoretical grounding for the informal notation we introduced when manipulating  $\delta$  as if it were a function of time.

**Proof** Consult Chapter 2.4 of [39].  $\square$

This result tells us that at every point of continuity, an approximation to the identity gives us exactly what we want—a way to precisely approximate the convolution behavior of the Dirac delta distribution using *signals* that are functions of time. Now, we study how approximations to the identity interact with impulse response. For simplicity, we will focus on the time-invariant case.

**Definition 2.38 (Approximate LTI Impulse Response Map)** Consider a continuous-time LTI system representation  $(A, B, C, D)$  and an approximation to the identity  $\{\delta_\epsilon\}_{\epsilon>0}$  consisting of piecewise continuous functions. The approximate LTI impulse response map with respect to  $\{\delta_\epsilon\}_{\epsilon>0}$  is the map  $H_\epsilon(\cdot) : \mathbb{R} \rightarrow \mathbb{R}^{p \times m}$ , parameterized by  $\epsilon > 0$  and defined,

$$H_\epsilon(t) = \begin{cases} C \exp(At)B + D\delta_\epsilon(t) & t \geq 0 \\ 0 & t < 0. \end{cases} \quad (2.326)$$

*Remark 2.48* Unlike in the previous subsection, where the continuous-time impulse response map was *not* a map we could evaluate as a function of time (due to the presence of the Dirac delta distribution),  $H_\epsilon(\cdot)$  is a map we can evaluate as a function of time. This is because  $\delta_\epsilon(\cdot)$  is a piecewise continuous function of time, and is therefore a true signal.

Using the approximate LTI impulse response map, we state an approximate version of Corollary 2.3 that relies on approximations to the identity.

**Proposition 2.25 (LTI Response via Approximation)** Consider a continuous-time, LTI system representation  $(A, B, C, D)$  with I/O map  $\rho$ . Let  $\{\delta_\epsilon\}_{\epsilon>0}$  be an approximation to the identity consisting of  $\delta_\epsilon \in PC(\mathbb{R}_{\geq 0}, \mathbb{R})$ , and  $H_\epsilon(\cdot)$  the approximate LTI impulse response map with respect to  $\{\delta_\epsilon\}_{\epsilon>0}$ . For any  $u(\cdot) \in PC(\mathbb{R}_{\geq 0}, \mathbb{R}^m)$ , and any  $t \in \mathbb{R}$  at which  $u(\cdot)$  is continuous,

$$\rho(t, 0, 0, u(\cdot)) = \lim_{\epsilon \rightarrow 0} \int_{\mathbb{R}} H_\epsilon(t - \tau) u(\tau) d\tau = \lim_{\epsilon \rightarrow 0} (H_\epsilon * u)(t). \quad (2.327)$$

*Remark 2.49* As a consequence of how we defined the approximate LTI impulse response map, we restrict ourselves to approximations defined on  $\mathbb{R}_{\geq 0}$  in this proposition.

**Proof** Consider a piecewise continuous input signal  $u(\cdot) \in PC(\mathbb{R}_{\geq 0}, \mathbb{R}^m)$ . If  $\mathbb{1}_{t \geq \tau}(\tau)$  represents the indicator function of  $t \geq \tau$ , the convolution  $H_\epsilon * u$  is computed at time  $t$  as,

$$(H_\epsilon * u)(t) = \int_{\mathbb{R}} H_\epsilon(t - \tau) u(\tau) d\tau \quad (2.328)$$

$$= \int_{\mathbb{R}} \mathbb{1}_{t \geq \tau}(\tau) [C \exp(A(t - \tau))B + D\delta_\epsilon(t - \tau)] u(\tau) d\tau \quad (2.329)$$

$$= \int_{\mathbb{R}} \mathbb{1}_{t \geq \tau}(\tau) C \exp(A(t - \tau))B u(\tau) d\tau + D \int_{\mathbb{R}} \delta_\epsilon(t - \tau) u(\tau) d\tau, \quad (2.330)$$

where we drop the indicator on the  $D$  integral under the assumption that each function  $\delta_\epsilon(\cdot) \in PC(\mathbb{R}_{\geq 0}, \mathbb{R})$ . Now, suppose  $t$  is a time at which  $u(\cdot)$  is continuous. Applying Theorem 2.15, it follows that, as  $\epsilon \rightarrow 0$ ,

$$\lim_{\epsilon \rightarrow 0} (H_\epsilon * u)(t) = \int_{\mathbb{R}} \mathbf{1}_{t \geq \tau}(\tau) C \exp(A(t - \tau)) B u(\tau) d\tau + \lim_{\epsilon \rightarrow 0} D \int_{\mathbb{R}} \delta_\epsilon(t - \tau) u(\tau) d\tau \quad (2.331)$$

$$= \int_{\mathbb{R}} \mathbf{1}_{t \geq \tau}(\tau) C \exp(A(t - \tau)) B u(\tau) d\tau + Du(t) \quad (2.332)$$

$$= C \int_0^t \exp(A(t - \tau)) B u(\tau) d\tau + Du(t) \quad (2.333)$$

$$= \rho(t, 0, 0, u(\cdot)), \quad (2.334)$$

where in the final step, we recognize the formula for the zero-state response.  $\square$

Thus, we can gain a similar result to Corollary 2.3 *without* the use of distributions—all we needed to state the proposition above was a family of well-behaved, continuous-time signals.

#### 2.4.4 The Laplace Transform

Let's summarize what we've done so far in this section. First, we determined that the response of a linear I/O system to any (sufficiently regular) input signal can be written in terms of the impulse response map. Following this, we showed that, for linear, time-invariant systems, the output is equal to the *convolution* of the impulse response and input signal. What's next in our study of I/O systems?

While the convolution sums and integrals we defined certainly provide clean expressions for the output of a system, they *are* still composed of sums and integrals that might be challenging to analyze directly. Can we do better than the convolution formulas? Let's set the stage by laying out a “wish list” of what we'd like in our analysis.

1. State equations become algebraic: state equations are generally challenging to work with. If we can reduce the state equations (which are ordinarily differential equations or recurrence relations) to purely *algebraic* equations, perhaps we'll be able to gain more insight into our systems.
2. Simplify convolutions: convolutions are also challenging operations to work with. We'd like a way to simplify convolution to an easy to compute operation.
3. Analyze long-term behavior: we'd like a way to study the long-term behavior of our system without performing hard analysis.
4. Simple interpretation: we should be able to determine important features of our systems by inspection.

Although this list seems a little ambitious, it turns out that in both the continuous and discrete-time LTI cases, we'll be able to meet every single point through the use of *transforms*. An amazing and quite surprising insight into this problem is that by transforming our objects of study into the complex plane, we can simplify the study of I/O systems to the study of complex, rational functions. What's more, under these transformations, convolution simplifies to *multiplication*.

We'll begin our study of these “magic transforms” in continuous-time, where we'll focus on the *Laplace transform*, and will then move to discrete-time, where we'll focus on the *Z-transform*. Here, we'll just touch upon the most basic aspects of these transforms—we'll develop some more nuanced aspects of transform theory as the course proceeds.

#### 2.4.4.1 A Little Complex Analysis

Above, we mentioned that our approach to achieving the points of our wish list will involve *transforming* system objects into objects in the complex plane. As such, it behooves us to learn the basic language of complex analysis, the study of functions in the complex plane.

First, we'll state some basic notation. Typically, when referring to a complex number, we'll use either  $s$  or  $z \in \mathbb{C}$ . Every complex number  $s \in \mathbb{C}$  admits a unique decomposition,

$$s = \operatorname{Re}(s) + j \operatorname{Im}(s), \quad (2.335)$$

where  $\operatorname{Re}(s) \in \mathbb{R}$  is the *real part* of  $s$ ,  $\operatorname{Im}(s) \in \mathbb{R}$  is the *complex part* of  $s$ , and  $j = \sqrt{-1}$ .<sup>10</sup> We stress—both the real *and* the imaginary part of a complex number are *real numbers*. The complex magnitude is a map  $|\cdot| : \mathbb{C} \rightarrow \mathbb{R}$ , taking,

$$s \mapsto |s| = \sqrt{(\operatorname{Re}(s))^2 + (\operatorname{Im}(s))^2}. \quad (2.336)$$

Since both  $\operatorname{Re}(s)$  and  $\operatorname{Im}(s)$  are real for all  $s \in \mathbb{C}$ , the complex magnitude is a *real-valued* function. One may show that the complex magnitude defines a *norm* on the complex numbers, which makes  $(\mathbb{C}, |\cdot|)$  a normed vector space over the field  $\mathbb{C}$ .

Since  $(\mathbb{C}, |\cdot|)$  has the structure of a normed vector space, all of the analysis tools we've developed for normed vector spaces immediately apply. Thus, using the complex magnitude, one can develop a notion of open and closed sets in  $\mathbb{C}$ , just like one might in  $\mathbb{R}$ . In particular, a set  $\Omega \subseteq \mathbb{C}$  is declared *open* if, for all  $s_0 \in \Omega$ , there exists an  $\epsilon > 0$  such that,

$$B_\epsilon(s_0) = \{s \in \mathbb{C} : |s - s_0| < \epsilon\} \subseteq \Omega. \quad (2.337)$$

Likewise, a set  $\Omega \subseteq \mathbb{C}$  is declared *closed* if its complement  $\Omega^c$  is open, and bounded if there exists an  $M \geq 0$  such that  $|s| \leq M$  for all  $s \in \Omega$ . A set  $\Omega$  is *compact* if and only if it is closed and bounded. Using these definitions, one may state a standard  $\epsilon$ - $\delta$  definition for continuity of complex functions, identical to that in a normed vector space.

An important function on  $\mathbb{C}$  is the *complex exponential*,  $\exp : \mathbb{C} \rightarrow \mathbb{C}$ , which is defined,

$$e^s = e^{\operatorname{Re}(s)} e^{j \operatorname{Im}(s)} := e^{\operatorname{Re}(s)} [\cos(\operatorname{Im}(s)) + j \sin(\operatorname{Im}(s))], \quad \forall s \in \mathbb{C}, \quad (2.338)$$

where  $e^{(\cdot)}$ ,  $\sin(\cdot)$ , and  $\cos(\cdot)$  are the real exponential, sine, and cosine functions. It's important to note: this is the *definition* of the exponential, not a theorem! The exponential gives us another way to write a complex number. Any  $s \in \mathbb{C}$  can be written in *polar form* as,

$$s = r e^{j\theta}, \quad r \geq 0, \theta \in [0, 2\pi). \quad (2.339)$$

<sup>10</sup> To quote [22] in reference to the use of  $j$  for  $\sqrt{-1}$ , “Only electrical engineers, and those under their influence, use this crazy notation.” If you feel more comfortable using  $i = \sqrt{-1}$ , you're welcome to do so—we won't consider any applications in circuit theory in this text, so there is no danger of conflating  $i = \sqrt{-1}$  with electrical current  $i$ .

Note that, for  $|s| > 0$ , the polar coordinates  $(r, \theta)$  for a given complex number are unique (up to  $\theta$  modulo  $2\pi$ ). For  $|s| = 0$ , this is not the case, as  $\theta$  becomes nonunique. The number  $r$  is to be interpreted as the magnitude of  $s$ ,  $r = |s|$ , and  $\theta$  as the angle between the vector  $(\operatorname{Re}(s), \operatorname{Im}(s))$  and the axis  $(1, 0)$ . One may show that, for  $s$  with  $\operatorname{Re}(s) \neq 0$ ,  $\theta = \arctan(\operatorname{Im}(s)/\operatorname{Re}(s))$ .

Now that we've discussed the very basics of complex numbers, we can talk about *calculus* in the complex plane. Let  $\Omega \subseteq \mathbb{C}$  be an open set and  $f : \Omega \rightarrow \mathbb{C}$  be a complex function. The function  $f$  is said to be *differentiable* at  $s_0 \in \mathbb{C}$  if,

$$\lim_{s \rightarrow s_0} \frac{f(s) - f(s_0)}{s - s_0}, \quad (2.340)$$

exists. In this case, the limit is declared to be the *derivative* of  $f$  at  $s_0$ , denoted  $f'(s_0)$ . It's important to note—although this definition looks identical to the definition in  $\mathbb{R}$ , here, the limit is being taken in the complex plane. We give differentiable complex functions the following special name.

**Definition 2.39 (Analytic Function)** Let  $\Omega \subseteq \mathbb{C}$  be an open set. A function  $f : \Omega \rightarrow \mathbb{C}$  is *analytic* at  $s_0 \in \Omega$  if there exists an  $\epsilon > 0$  such that  $f$  is differentiable at every point in  $B_\epsilon(s_0)$ . If  $f$  is analytic at every  $s_0 \in \Omega$ , then  $f$  is said to be *analytic on  $\Omega$* .

*Remark 2.50* In the theory of single-variable complex functions the terms “analytic” and “holomorphic” are often used interchangeably. One might see an *analytic* function defined as a function with a convergent power series in a neighborhood of a point, and a *holomorphic* function being a (complex) differentiable function. For single-variable complex functions, these two definitions are *equivalent*, and can be used interchangeably. Since differentiability is an easier condition to check, we define analyticity via differentiability.

Amazingly, one may show that a function  $f : \Omega \rightarrow \mathbb{C}$  is analytic at  $s_0 \in \mathbb{C}$  if and only if it has a convergent power series expansion in a neighborhood of  $s_0$ . This is *not* true even for infinitely differentiable functions<sup>11</sup> in  $\mathbb{R}$ . These basic properties hint that complex analysis can often be more revealing than real analysis—transforming a real function into the complex plane immediately gives us access to these new results!

Before we get too carried away, it's important to perform a few sanity checks regarding the complex derivative. Since the complex derivative has essentially the same definition as the derivative in  $\mathbb{R}$  (with the limit being taken in  $\mathbb{C}$  instead of  $\mathbb{R}$ ), it enjoys all of the same basic properties (linearity, chain rule, product rule, etc.) as the real derivative.

**Theorem 2.16 (Properties of the Complex Derivative)** Let  $f, g : \Omega \rightarrow \mathbb{C}$  be analytic functions on an open set  $\Omega \subseteq \mathbb{C}$ .

1. *Linearity:*  $f + g$  is analytic in  $\Omega$  and  $(f + g)' = f' + g'$ .
2. *Product Rule:*  $fg$  is analytic in  $\Omega$  and  $(fg)' = f'g + fg'$ .
3. *Chain Rule:* If  $U \subseteq \mathbb{C}$  is open,  $h : U \rightarrow \mathbb{C}$  is analytic on  $U$ , and  $f(\Omega) \subseteq U$ , then  $h \circ f$  is analytic on  $\Omega$  with  $(h \circ f)'(s) = h'(f(s))f'(s)$ .

**Exercise 2.23** Write a version of Theorem 2.16 in which each function  $f, g, h$  is analytic at a point. Use Theorem 2.16 to prove your theorem.

We refer the reader to [38] for proofs of the results stated in this section, as well as for a more thorough treatment of complex analysis.

<sup>11</sup> A counterexample can be constructed by taking a nonzero, smooth function which is very *flat* in a region around the origin. The function  $f(x) = e^{-1/x}$ ,  $x < 0$ ,  $f(x) = 0$ ,  $x \leq 0$  is an example of this.



### 2.4.4.2 Definition & Basic Properties of the Laplace Transform

Now that we've reviewed some of the basic language of complex analysis, we're ready to describe transformations which take *real* functions and transform them into *complex* functions. In this section, we'll study the Laplace transform, which performs this transformation for certain classes of continuous-time signals. We'll first present the definition of the transform and then discuss its properties.

**Definition 2.40 (Laplace Transform)** Consider a piecewise continuous, matrix-valued signal,  $f(\cdot) \in PC(\mathbb{R}, \mathbb{R}^{n \times m})$ . The one-sided Laplace transform of  $f$  is the complex-valued function,  $\hat{F} : \Omega \rightarrow \mathbb{C}^{n \times m}$ , defined

$$\hat{F}(s) = \mathcal{L}(f)(s) := \lim_{\epsilon \rightarrow 0^-} \int_{\epsilon}^{\infty} e^{-st} f(t) dt = \int_{0^-}^{\infty} e^{-st} f(t) dt, \quad (2.341)$$

where  $\Omega := \{s \in \mathbb{C} : \mathcal{L}(f)(s) \text{ converges absolutely}\}$  is called the *region of absolute convergence* of  $\hat{F}$ .

*Remark 2.51* Here, we define the Laplace transform using an integral of a matrix-valued function. Recall that integrals of matrix-valued functions are defined *element-wise*—that is, we integrate each element of the matrix-valued function individually. In order to take the Laplace transform of a matrix-valued function, we therefore take the transform of each element function of the matrix.

*Remark 2.52* The notation  $0^-$  means that the lower bound of the integral should be taken in the limit as 0 is approached from the *left*. This lets the Laplace transform interact well with approximations to the identity, which approximate the Dirac delta. In particular, if we did not include this left limit, the Laplace transform would *not* correctly interact with approximations that are defined on all of  $\mathbb{R}$ . If we're given a signal  $f(\cdot)$  that is only defined on  $\mathbb{R}_{\geq 0}$ , we compute the limit as  $t \rightarrow 0^-$  under the assumption that  $f(t) = 0$  for  $t < 0$ —this aligns with our “standard” way of extending the domains of signals to all of  $\mathbb{R}$ .

*Remark 2.53* Typically, we will use a capital letter with a hat to denote the Laplace transform of a signal. For instance, if  $f$  is a signal, we will denote by  $\hat{F}$  its Laplace transform. In some circumstances, one will see the transform written without the hat—we'll use the hat here for the sake of clarity. This is useful when writing the transform of a signal which is already specified by a capital letter.

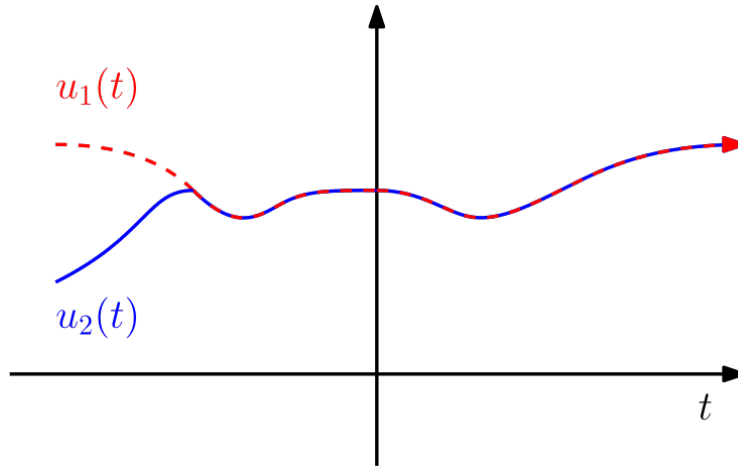
*Remark 2.54* By inspection, we can tell that, for an arbitrary signal, the Laplace transform is *not* guaranteed to converge for every  $s \in \mathbb{C}$ . For instance, if  $f$  grows faster than  $\text{Re}(e^{st})$ , the integral will diverge at  $s$ . This leads us to define  $\hat{F}$  on the region of absolute convergence,  $\Omega$ . We'll often refer to the “region of absolute convergence” as the “region of convergence,” and will use the letters R.O.C. for shorthand.

From this definition, we observe that a Laplace transform integrates  $f$  (which is a function of  $t$ ) against  $e^{-st}$  (which is function of  $t$  and  $s$ ), to get a function of  $s$ . Since the time variable is integrated out, *all that remains* when the transform is computed is a function of  $s$ . Just like we refer to  $t$  as a “time” variable, we will refer to  $s$  as *frequency* variable. For now, we will justify the use of this language via the complex exponential,

$$e^{st} = e^{\text{Re}(st)} (\cos(\text{Im}(st)) + j \sin(\text{Im}(st))). \quad (2.342)$$

This function, which appears in the Laplace transform, is defined in terms of sines and cosines whose frequencies are determined by  $s$ . We'll explore a deeper connection between  $s$  and frequency later in the course, when we discuss *frequency response*. In line with this language, we refer to the real numbers  $\mathbb{R}$  as the time domain and the complex numbers  $\mathbb{C}$  as the *frequency domain*. Thus, one says that the Laplace transform takes a signal from the *time domain* and transforms it into the *frequency domain*.

It's *extremely* important to note that the Laplace transform we've defined above is only *one-sided*. This means that the integral in the transform *does not* extend to  $t = -\infty$ . This one-sided definition reflects the fact that the LTI systems we're interested in analyzing are *causal*. Above, we showed that the impulse response map of a causal LTI system is zero for  $t < 0$ —starting the transform at  $t = 0^-$  is therefore justifiable.



**Fig. 2.7** The signals  $u_1$  and  $u_2$  share the same one-sided Laplace transform.

Above, we defined the *region of absolute convergence*,  $\Omega$ , to be the region in which the integral defining the transform converges absolutely. Which signals have a nonempty region of absolute convergence? Do there exist signals with no region of convergence? To determine the class of signals with a well-defined region of convergence, let's take another look at the definition of the Laplace transform. We have,

$$\mathcal{L}(f)(s) = \int_{0^-}^{\infty} e^{-st} f(t) dt. \quad (2.343)$$

In order for this integral to converge absolutely for some values of  $s$ , it seems appropriate to request that  $f$  be bounded by some exponential growth. This leads us to the following definition.

**Definition 2.41 (Function of Exponential Order)** A function  $f \in PC(\mathbb{R}_{\geq 0}, \mathbb{R}^{n \times m})$  is of exponential order  $\alpha$  if there exists an  $M \in \mathbb{R}_{\geq 0}$  for which  $\|f(t)\| \leq Me^{\alpha t}$ , for all  $t \in \mathbb{R}_{\geq 0}$ .

*Remark 2.55* Notice that the condition  $\|f(t)\| \leq Me^{\alpha t}$  is specified in terms of the values of the signal  $f$ , which are members of a finite-dimensional vector space. Due to norm equivalence in finite dimensional vector spaces, the choice of norm  $\|\cdot\|$  on  $\mathbb{R}^{n \times m}$  doesn't matter. A different choice of norm will simply result in a scaling of  $M$  by a positive constant.

Now, we show that all functions of exponential order have well-defined Laplace transforms on a nonempty region of convergence.

**Proposition 2.26 (Laplace Transform of a Function of Exponential Order)** *Let  $f \in PC(\mathbb{R}_{\geq 0}, \mathbb{R}^{n \times m})$  be a function of exponential order  $\alpha$ . Then, the region of absolute convergence of the Laplace transform of  $f$  contains the set*

$$\{s \in \mathbb{C} : \operatorname{Re}(s) > \alpha\} \subseteq \mathbb{C}. \quad (2.344)$$

**Proof** Suppose  $f$  satisfies  $\|f(t)\| \leq Me^{\alpha t}$  for all  $t \geq 0$ . Then, one has

$$\int_{0^-}^{\infty} |e^{-st}| \|f(t)\| dt \leq M \int_{0^-}^{\infty} e^{(\alpha - \operatorname{Re}(s))t} dt. \quad (2.345)$$

This integral will converge when  $\alpha - \operatorname{Re}(s) < 0$ , or when  $\alpha < \operatorname{Re}(s)$ . Thus, for all  $s$  satisfying  $\alpha < \operatorname{Re}(s)$ , the Laplace transform will converge absolutely. This implies the given set is contained in the region of absolute convergence of the Laplace transform of  $f$ .  $\square$

As an illustration of what can go wrong if we don't enforce the exponential order assumption, we consider a signal whose Laplace transform has an empty region of convergence.

*Example 2.4* To choose a signal with no Laplace transform, we choose a signal that grows faster than any exponential. For instance, consider the signal,  $f(t) = e^{e^t}$ , which grows faster than an exponential since  $e^t$  grows faster than  $t$ . Such a signal has no Laplace transform.

Now that we've defined the Laplace transform and a basic class of signals with a well-defined transform, we'll get some practice computing Laplace transforms.

**Theorem 2.17 (Common One-Sided Laplace Transforms)** *Consider the following collection of signals, transforms, and regions of absolute convergence of their transforms.*

| Signal Name   | Signal  | Laplace Transform                   | Region of Convergence                         |
|---------------|---|-------------------------------------|---|
| Dirac Delta   | $\delta(t)$   | 1                                   | $\mathbb{C}$                                  |
| Unit Step     | $\mathbf{1}(t) = \begin{cases} 1 & t \geq 0 \\ 0 & t < 0 \end{cases}$ | $\frac{1}{s}$                       | $\operatorname{Re}(s) > 0$                    |
| Unit Ramp     | $t$   | $\frac{1}{s^2}$                     | $\operatorname{Re}(s) > 0$                    |
| Polynomial    | $t^n, n \in \mathbb{N}$   | $\frac{n!}{s^{n+1}}$                | $\operatorname{Re}(s) > 0$                    |
| Exponential   | $e^{at}$  | $\frac{1}{s-a}$                     | $\operatorname{Re}(s) > \operatorname{Re}(a)$ |
| Sine          | $\sin(\omega t)$  | $\frac{\omega}{s^2 + \omega^2}$     | $\operatorname{Re}(s) > 0$                    |
| Cosine        | $\cos(\omega t)$  | $\frac{s}{s^2 + \omega^2}$          | $\operatorname{Re}(s) > 0$                    |
| Damped Sine   | $e^{at} \sin(\omega t)$   | $\frac{\omega}{(s-a)^2 + \omega^2}$ | $\operatorname{Re}(s) > \operatorname{Re}(a)$ |
| Damped Cosine | $e^{at} \cos(\omega t)$   | $\frac{s-a}{(s-a)^2 + \omega^2}$    | $\operatorname{Re}(s) > \operatorname{Re}(a)$ |

Shortly, we'll see that the Laplace transform behaves well under algebraic combinations of signals—this makes it possible to derive the transforms of more complex signals from the table above. In order to illustrate the general technique of computing a Laplace transform, we'll derive the transform of the Dirac delta and the unit step functions, arguably the two most important transforms signals. The rest of the table entries are found by grinding out integrals.

*Example 2.5 (Laplace Transform of the Dirac Delta)* Let's work out the Laplace transform of the Dirac delta function. Using our informal integral “rules,” we have

$$\mathcal{L}(\delta)(s) = \int_{0^-}^{\infty} e^{-st} \delta(t) dt = \int_{-\infty}^{\infty} e^{s \cdot 0 - st} \delta(t) dt = e^{st}|_{t=0} = 1, \quad (2.346)$$

where we conclude that the lower bound can be extended to  $-\infty$ , since the Dirac delta “function of time” vanishes for  $t < 0$ . This derivation is made rigorous using approximations to the identity (see the previous, starred subsection for the details).

*Example 2.6 (Laplace Transform of the Unit Step)* Next, we calculate the transform of the unit step signal,  $\mathbf{1}(\cdot)$ . We have,

$$\mathcal{L}(\mathbf{1})(s) = \int_0^{\infty} e^{-st} dt = \left[ -\frac{1}{s} e^{-st} \right]_0^{\infty} = \frac{1}{s} \text{ (for } \operatorname{Re}(s) > 0). \quad (2.347)$$

**Exercise 2.24** Complete the rest of the Laplace transform table in Theorem 2.17. See [21] (or any other standard text on signals and systems) for a solution.

Let's examine the table of Laplace transforms and their regions of convergence in a more critical light. Above, we showed that the Laplace transform of the unit step function is,

$$\frac{1}{s} \text{ with R.O.C. } \operatorname{Re}(s) > 0. \quad (2.348)$$

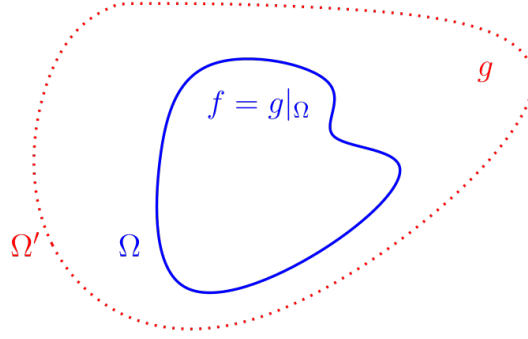
The function  $1/s$  is defined *everywhere* in the complex plane but  $s = 0$ —yet, the region of convergence of  $\mathcal{L}(\mathbf{1})(s)$  is only  $\operatorname{Re}(s) > 0$ . This seems a little bit contradictory! Let's look at another example. The transform of  $e^{at}$  is  $1/(s - a)$ —defined everywhere but  $s = a$ —yet the transform converges absolutely for  $\operatorname{Re}(s) > a$ .

It appears as if we *should* be able to define our transforms on much larger regions than the region of convergence of the integral. Can we extend the domain of definition of our transforms in some canonical way? An important concept from complex analysis, *analytic continuation*, lets us do exactly this.

**Theorem 2.18 (Analytic Continuation)** Consider two complex functions,  $f : \Omega \rightarrow \mathbb{C}$  and  $g : \Omega' \rightarrow \mathbb{C}$  on nonempty open domains  $\Omega \subseteq \Omega' \subseteq \mathbb{C}$ . If  $f$  is analytic on  $\Omega$ ,  $g$  is analytic on  $\Omega'$ , and  $f(s) = g(s)$  for all  $s \in \Omega$ , then  $g$  is the unique analytic function on  $\Omega'$  for which  $f = g|_{\Omega}$ . In this case,  $g$  is called the analytic continuation of  $f$  to  $\Omega'$ .

**Proof** See Chapter 2, Corollary 4.9 of [38]. □

Let's think about the implications of analytic continuation for the transforms we discussed above. Consider, for example, the transform of  $\mathbf{1}(t)$ ,  $\mathbf{1}(s) = 1/s$ , which has region of convergence  $\Omega = \operatorname{Re}(s) > 0$  and is analytic (complex-differentiable) on  $\Omega$ . We know that  $1/s$  can be defined on the open set  $\Omega' = \mathbb{C} \setminus \{0\}$ , and is also analytic on  $\Omega'$ . Thus, by Theorem



**Fig. 2.8** If two analytic functions agree on a small domain, then there is only *one* way to extend the analytic function on the smaller domain to the larger domain. Here, this means that  $g$  is the *unique* analytic continuation of  $f$  to  $\Omega'$ —the only possible analytic function on  $\Omega'$  that matches  $f$  on  $\Omega$ .

2.18, the *unique analytic continuation* of  $1/s$  (on  $\text{Re}(s) > 0$ ) to  $\mathbb{C} \setminus \{0\}$  is given by  $1/s$  on  $\Omega' = \mathbb{C} \setminus \{0\}$ . Thus, provided there exists an analytic continuation of the transform on a domain larger than the region of convergence, we can *unambiguously* extend the domain of the transform. This is because the analytic continuation is the *only possible* analytic extension of the original transform, so there is no ambiguity in “choosing between” different possible analytic extensions.

Due to the uniqueness of analytic continuation, when the extension of a transform is easy to compute, we can easily identify a transform with its analytic continuation to a larger domain. This lets us work with transforms such as  $1/s$  on a much larger domain than just the region of convergence of the integral. To illustrate this point further, suppose we have a signal with transform,

$$F(s) = \frac{1}{(s-a)(s+b)}, \quad (2.349)$$

on some nonempty region of convergence  $\Omega$ . We observe by inspection that we can analytically continue the transform from its region of convergence  $\Omega$  to  $\mathbb{C} \setminus \{a, -b\}$ . Thus, we can unambiguously work with the transform on this larger domain. It's important to note—for general signals that have a Laplace transform, extensions are *not* as obvious as in the cases above. However, as a large class of signals of interest have simple transforms, this technique is something we can often apply.

Now that we've discussed some basic examples of Laplace transforms, we can study some of their important properties. First, we'll state the three *most fundamental* properties of the Laplace transform. Following this, we'll examine some more general-purpose properties of the transform.

**Theorem 2.19 (Key Properties of the Laplace Transform)** *Let  $f, g \in PC(\mathbb{R}_{\geq 0}, \mathbb{R})$  be signals of exponential orders  $\alpha$  and  $\beta$ , respectively.*

1. *Analyticity:* For any  $\epsilon > 0$ , at all  $s$  for which  $\text{Re}(s) > \alpha + \epsilon$ ,  $\mathcal{L}(f)(s)$  is analytic.
2. *Linearity:*  $\mathcal{L}(k_1 f + k_2 g)(s) = k_1 \mathcal{L}(f)(s) + k_2 \mathcal{L}(g)(s) \quad \forall k_1, k_2 \in \mathbb{R}, s : \text{Re}(s) > \max\{\alpha, \beta\}$ .
3. *Convolution:*  $\mathcal{L}(f * g)(s) = \mathcal{L}(f)(s) \cdot \mathcal{L}(g)(s)$  for all  $s$  satisfying  $\text{Re}(s) > \max\{\alpha, \beta\}$ .

*Remark 2.56* Remember: if  $\mathcal{L}(f)$  and  $\mathcal{L}(g)$  have analytic continuations to larger domains, we can extend the domains on which these properties hold!

*Remark 2.57* Each of these properties is easily extended to the case where  $f$  and  $g$  are vector or matrix-valued signals! Here, one simply needs to take care to ensure the dimensions of  $f$  and  $g$  match.

**Proof (Sketch)** We'll give a proof sketch of these three properties—for a formal proof, one should be more precise in dealing with the convergence of indefinite integrals. We start with property (1). Fix  $\epsilon > 0$ , and consider  $\operatorname{Re}(s) > \alpha + \epsilon$ . Since exponential growth dominates polynomial growth, it follows that, if  $|f(t)| \leq me^{\alpha t} \forall t \geq 0$ , there exists an  $m \geq 0$  for which  $|tf(t)| \leq m'e^{(\alpha+\epsilon)t} \forall t \geq 0$  (we'll prove this formally in the next chapter). Then,

$$\frac{d}{ds} \int_{0-}^{\infty} e^{-st} f(t) dt = \int_{0-}^{\infty} \frac{d}{ds} e^{-st} f(t) dt \quad (2.350)$$

$$= \int_{0-}^{\infty} -te^{-st} f(t) dt \quad (2.351)$$

$$= -\mathcal{L}(tf(t)), \quad (2.352)$$

where the integral converges for all  $s$  satisfying  $\operatorname{Re}(s) > \alpha + \epsilon$ . Thus, the derivative of the Laplace transform exists for  $\operatorname{Re}(s) > \alpha + \epsilon$ . We conclude the analyticity of the transform. Next, we prove the linearity property. We have, for  $k_1, k_2 \in \mathbb{R}$ , and  $\operatorname{Re}(s) > \max\{\alpha, \beta\}$

$$\mathcal{L}(k_1 f + k_2 g) = \int_{0-}^{\infty} e^{-st} (k_1 f(t) + k_2 g(t)) dt \quad (2.353)$$

$$= k_1 \int_{0-}^{\infty} e^{-st} f(t) dt + k_2 \int_{0-}^{\infty} e^{-st} g(t) dt \quad (2.354)$$

$$= k_1 \mathcal{L}(f)(s) + k_2 \mathcal{L}(g)(s), \quad (2.355)$$

where we use that each integral converges absolutely for  $\operatorname{Re}(s) > \max\{\alpha, \beta\}$ . This shows the linearity property. Finally, we show the convolution property. We have,

$$\mathcal{L}(f * g)(s) = \int_{0-}^{\infty} e^{-st} \int_{0-}^{\infty} f(t - \tau) g(\tau) d\tau dt \quad (2.356)$$

$$= \int_{0-}^{\infty} \int_{0-}^{\infty} e^{-st} f(t - \tau) g(\tau) d\tau dt \quad (2.357)$$

$$= \int_{0-}^{\infty} \int_{0-}^{\infty} e^{-s(t-\tau)} e^{-s\tau} f(t - \tau) g(\tau) d\tau dt \quad (2.358)$$

$$= \int_{0-}^{\infty} \int_{0-}^{\infty} e^{-s(t-\tau)} f(t - \tau) e^{-s\tau} g(\tau) d\tau dt \quad (2.359)$$

$$= \int_{0-}^{\infty} e^{-s\tau} g(\tau) \int_{0-}^{\infty} e^{-s(t-\tau)} f(t - \tau) dt d\tau \quad (2.360)$$

$$= \int_{0-}^{\infty} e^{-s\tau} g(\tau) \int_{\tau}^{\infty} e^{-s(t-\tau)} f(t - \tau) dt d\tau, \quad (2.361)$$

where in the last step, we used that the signal  $f$  is zero for  $t < \tau$ . We recognize the inner integral as the Laplace transform of  $f$ , which converges for  $\operatorname{Re}(s) > \max\{\alpha, \beta\}$ . Thus,

$$= \int_{0^-}^{\infty} e^{-s\tau} g(\tau) \mathcal{L}(f)(s) d\tau \quad (2.362)$$

$$= \mathcal{L}(f)(s) \int_{0^-}^{\infty} e^{-s\tau} g(\tau) d\tau \quad (2.363)$$

$$= \mathcal{L}(f)(s) \cdot \mathcal{L}(g)(s), \quad (2.364)$$

where the final transform also converges for  $\operatorname{Re}(s) > \max\{\alpha, \beta\}$ . We conclude that one-sided convolution becomes multiplication under the one-sided Laplace transform!  $\square$

**Exercise 2.25** Restate Theorem 2.19 in the case where  $f$  and  $g$  are matrix-valued signals.

This result highlights some of the principal reasons why the Laplace transform will be so useful for us. First, the Laplace transform is *linear* in signals—this means that, for any of the simple signals we stated transforms for, we *immediately* know the transforms of their linear combinations. Secondly, we found that the Laplace transform of the convolution of two signals becomes the product of the Laplace transforms of the signals. Thus, the Laplace transform fulfills our wish of simplifying the convolution operation. This is key for studying the response of linear, time-invariant systems, which is governed by convolution.

Now that we've stated these three fundamental properties, we summarize a number of additional properties concerning the interaction of the Laplace transform with integrals, derivatives, and delays.

**Theorem 2.20 (Further Properties of the Laplace Transform)** *Let  $f : \mathbb{R} \rightarrow \mathbb{R}$  be a signal of exponential order  $\alpha$  on  $\mathbb{R}_{\geq 0}$  with one-sided Laplace transform  $F(s)$  on  $\Omega$ . Then, the following properties hold on a nonempty subset of  $\Omega$ .*

| Operation                               | Signal                           | Laplace Transform                                    |
|---|----------------------------------|--|
| Derivative                              | $f'(t)$                          | $sF(s) - f(0^-)$                                     |
| $n$ 'th Derivative                      | $f^{(n)}(t)$                     | $s^n F(s) - s^{n-1} f(0^-) - \dots - f^{(n-1)}(0^-)$ |
| Integral                                | $\int_{0^-}^{t^+} f(\tau) d\tau$ | $\frac{F(s)}{s}$                                     |
| Product with $t$                        | $tf(t)$                          | $-F'(s)$   |
| Product with $t^n$ , $n \in \mathbb{N}$ | $t^n$                            | $(-1)^n F^{(n)}(s)$                                  |
| Delay by $\tau$                         | $f(t - \tau)$                    | $e^{-\tau s} F(s)$                                   |
| Time scaling by $a \neq 0$              | $f(at)$                          | $\frac{1}{ a } F(s/a)$                               |

**Exercise 2.26** Prove Theorem 2.20 via direct calculation. See [21] (or any other standard text on signals and systems) for a solution.

As a nice application of Theorem 2.20, one can confirm some of the basic Laplace transforms stated earlier using the various operations above. For instance, the ramp function is the integral of the step function, which is consistent with the ramp function having transform  $1/s^2$  and the step function having transform  $1/s$ .

### 2.4.4.3 Transfer Functions

Now that we've studied the basic definition and properties of the Laplace transform, we're ready to apply it to the input/output analysis of continuous-time, LTI systems. Earlier, we showed that for an input signal  $u(\cdot) \in PC(\mathbb{R}_{\geq 0}, \mathbb{R}^m)$ , the zero-state response of a continuous-time LTI system,  $y(t)$ , is computed

$$y(t) = (H * u)(t), \quad (2.365)$$

at every  $t \in \mathbb{R}$  at which  $u$  is continuous. Above, we showed that one-sided convolution becomes multiplication under the Laplace transform. Therefore, if we wish to understand the zero-state response of a system using the Laplace transform, we can apply the rule,

$$\mathcal{L}(y)(s) = \mathcal{L}(H * u)(s) = \mathcal{L}(H)(s)\mathcal{L}(u)(s). \quad (2.366)$$

Since convolution becomes multiplication in the frequency domain, *all we need* to compute the transform of the zero-state response is the product of the transform of the impulse response and the transform of the input signal. Thus, in the frequency domain, multiplication by the transform of the impulse response directly transfers us from input to output. This yields the following definition.

**Definition 2.42 (Transfer Function)** Consider a continuous-time LTI system with LTI impulse response map  $H(\cdot) : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^{p \times m}$ . The transfer function of the system is the map  $\hat{H} : \Omega \subseteq \mathbb{C} \rightarrow \mathbb{C}^{p \times m}$ , defined

$$\hat{H}(s) = \mathcal{L}(H)(s), \quad \forall s \in \Omega, \quad (2.367)$$

where  $\Omega$  is the region of convergence of the transform.

*Remark 2.58* Typically, one uses either  $\hat{H}$  or  $\hat{G}$  to refer to the transfer function of a system. We'll interchange between the two.

*Remark 2.59* Remember,  $\hat{H}$  can be extended to a larger domain than  $\Omega$  if it has an analytic continuation to a larger domain! We'll see a few examples of this shortly.

*Remark 2.60* Since  $\hat{H}(s) \in \mathbb{C}^{p \times m}$  for all  $s \in \Omega$ , it follows that a MIMO system will have a matrix-valued transfer function, while a SISO system will have a scalar-valued transfer function. Since one has  $Y(s) = \hat{H}(s)U(s)$ , we can calculate the transfer function via  $Y(s)/U(s) = \hat{H}(s)$  in the SISO case.

A few questions immediately appear upon making this definition. Does the Laplace transform of an impulse response always have a nonempty region of convergence? How does one most easily compute the Laplace transform of  $H(\cdot)$ ? Let's start with the first question. Above, we showed that every map of exponential order has a nonempty region of convergence. Is the impulse response map of an LTI system a map of exponential order? Recall that the LTI impulse response map is defined,

$$H(t) = \begin{cases} C \exp(At)B + D\delta & t \geq 0 \\ 0 & t < 0. \end{cases} \quad (2.368)$$



By linearity of the Laplace transform it's sufficient for  $C \exp(At)B$  and  $D\delta(t)$  to have well-defined transforms in order for  $H(\cdot)$  to have a well-defined transform. We know that  $D\delta$  has a transform  $\mathcal{L}(D\delta)(s) = D$ , since  $\mathcal{L}(\delta)(s) = 1$  for all  $s \in \mathbb{C}$ . What about  $C \exp(At)B$ ? It certainly *seems* that the matrix exponential should be of exponential order! Fortunately, this turns out to be true—we'll accept this as a fact for now and will return to the proof next chapter, in our study of stability.

**Fact** For any  $A \in \mathbb{R}^{n \times n}$ , the matrix exponential  $\exp(At)$  is of exponential order.  $\square$

Since  $\exp(At)$  is of exponential order, we conclude that  $C \exp(At)B$  is also of exponential order. The LTI impulse response map  $H(\cdot)$  will therefore have a well-defined Laplace transform on a nonempty region of convergence.

Now, we focus on to the second question: how do we actually *compute* the transfer function of a given LTI system? One option is to compute the transform directly from the definition of  $H(t)$ , posed in terms of the matrix exponential. However, since the whole point of the Laplace transform is to avoid dealing with things like the matrix exponential, it would hardly be a great success if we needed to compute a Jordan form, change coordinates, and compute the transform of each entry. We'll show that there is a *much simpler* way to compute the transfer function. What's more, this simpler method will give us a way to compute the matrix exponential *without* appealing to the Jordan form. First, consider the following lemma.

**Lemma 2.14 (Characterizing the Transfer Function)** *Consider a continuous-time, LTI system representation  $(A, B, C, D)$  with LTI impulse response map  $H$ . A function  $\hat{H} : \Omega \rightarrow \mathbb{C}^{p \times m}$  is the transfer function of the system if and only if  $\hat{Y}(s) = \hat{H}(s)\hat{U}(s)$  for all transforms  $\hat{U}(s)$  and  $\hat{Y}(s)$  of admissible input signals  $u(\cdot)$  and their corresponding zero-state responses  $y(\cdot)$ .*

**Proof** First, suppose  $\hat{H}$  is the transfer function of the system. Then, by the convolution property,  $\hat{Y}(s) = \hat{H}(s)\hat{U}(s)$ . Now, we show the other direction. Suppose a function  $\hat{H}$  satisfies  $\hat{Y}(s) = \hat{H}(s)\hat{U}(s)$  for all admissible inputs and their corresponding zero-state outputs. Consider the input signal  $u = \delta e_j$ , where  $e_j \in \mathbb{R}^m$  is the  $j$ 'th standard basis vector of  $\mathbb{R}^m$ . Then,  $\hat{Y}(s) = \hat{H}(s)e_j$ . By definition of the impulse response map, however, the  $j$ 'th column of  $\mathcal{L}(H)(s)$  must also be  $\hat{H}(s)e_j$ . We conclude that  $\hat{H}$  must be the transfer function of the system.  $\square$

With this lemma in mind, we show how the transfer function of a continuous-time, LTI system is easily computed.

**Proposition 2.27 (Transfer Function of an LTI System Representation)** *Consider a continuous-time, LTI system representation  $(A, B, C, D)$ . The transfer function  $\hat{H}$  of the system is computed,*

$$\hat{H}(s) = C(sI - A)^{-1}B + D, \quad \forall s \in \Omega, \quad (2.369)$$

where  $\Omega = \mathbb{C} \setminus \text{spec}(A)$ , the complex plane minus the eigenvalues of  $A$ .

**Proof** Recall that the system representation  $(A, B, C, D)$  satisfies,

$$\dot{x}(t) = Ax(t) + Bu(t) \quad (2.370)$$

$$y(t) = Cx(t) + Du(t). \quad (2.371)$$

Let's take the Laplace transform of both of these equations for zero initial condition,  $x(0) = 0$ , and admissible input  $u$ . Computing the transforms, one has,

$$s\hat{X}(s) = A\hat{X}(s) + B\hat{U}(s) \quad (2.372)$$

$$\hat{Y}(s) = C\hat{X}(s) + D\hat{U}(s). \quad (2.373)$$

Since  $(sI - A)$  is nonsingular for  $s \notin \text{spec}(A)$ , for  $s \in \mathbb{C} \setminus \text{spec}(A)$ ,  $\hat{X}$  must satisfy,

$$\hat{X}(s) = (sI - A)^{-1}B\hat{U}(s). \quad (2.374)$$

Substituting into the formula for  $Y$ , it follows that,

$$\hat{Y}(s) = (C(sI - A)^{-1}B + D)\hat{U}(s). \quad (2.375)$$

We conclude that,

$$\hat{H}(s) = C(sI - A)^{-1}B + D. \quad (2.376)$$

The function  $(sI - A)^{-1}$  is defined (and analytic) on  $\mathbb{C} \setminus \text{spec}(A)$ . Identifying the transfer function of the system with its analytic continuation, the result follows.  $\square$

This gives us an easy way to compute the transfer function of any continuous-time, LTI system representation—all we need to do is invert a single matrix function of  $s$  (computers are quite good at this for small matrices) and perform some matrix multiplication. Now that we have a nice way to compute transfer functions, we can use them to solve for the zero-state response of a system in the time domain. In order to convert from the frequency domain to the time domain, we define the inverse Laplace transform.

**Definition 2.43 (Inverse Laplace Transform)** The inverse Laplace transform is a map  $\mathcal{L}^{-1}$  taking a complex-valued function  $\hat{H} : \Omega \rightarrow \mathbb{C}^{p \times m}$  to a real-valued signal  $H(\cdot) : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^{p \times m}$  satisfying  $\mathcal{L}(H)(s) = \hat{H}(s)$  for all  $s \in \Omega$ .

*Remark 2.61* There are a few subtleties regarding the definition of the inverse Laplace transform. First, the inverse Laplace transform is *not* guaranteed to exist for any complex-valued function! Second, if the inverse Laplace transform *does* exist, it is not guaranteed to be unique. For a general complex-valued function, the best one can hope for is a unique inverse Laplace transform up to equality outside of sets of measure zero—sets which are ignored by integrals. This is a consequence of the Laplace transform being an integral transform. If one takes signals  $f, g \in PC(\mathbb{R}_{\geq 0}, \mathbb{R}^{p \times m})$  which are equal outside of a set of measure zero, the signals will have the same Laplace transform. Therefore, the inverse Laplace transform of  $\mathcal{L}(f) = \mathcal{L}(g)$  can at best be uniquely defined up to equality outside of sets of measure zero.

The existence of an inverse Laplace transform is generally challenging to establish for an arbitrary, complex-valued function. However, for a large class of functions, a well-defined, *unique* inverse does exist.

**Lemma 2.15 (Inverse Laplace Transform of a Strictly Proper Rational Function)** Consider a complex function which is the ratio of two polynomials with real coefficients,

$$\hat{G}(s) = \frac{a_m s^m + \dots + a_1 s + a_0}{b_n s^n + \dots + b_1 s + b_0}, \quad a_i, b_i \in \mathbb{R}, \quad a_m, b_n \neq 0. \quad (2.377)$$

If  $m < n$ , then  $\hat{G}$  is called a strictly proper rational function. If  $m < n$ , then  $\hat{G}$  has a unique inverse Laplace transform, up to equality outside of sets of measure zero.

*Remark 2.62* Loosely speaking, time-domain signals comprised of algebraic combinations of polynomials, trigonometric functions, and exponential functions of time will have transforms of the form above. For such functions, inverse Laplace transforms are uniquely defined. We refer the interested reader to [22], Chapter 15.3, for the details.

*Remark 2.63* Recalling that the Laplace transform of a matrix-valued signal is defined *element-wise*, we can extend this result to the case where each element of a matrix-valued function is a strictly proper rational function.

**Proof (Sketch)** We'll provide a sketch of the existence proof in the case where the roots of the numerator and denominator polynomials are real and non-repeated. For the general case and the details of uniqueness, we refer the reader to [22]. If the numerator and denominator polynomials of  $\hat{G}$  have real, non-repeated roots,  $\hat{G}$  can be factored,

$$\hat{G}(s) = \frac{(s - s_{a,1}) \cdots (s - s_{a,m})}{(s - s_{b,1}) \cdots (s - s_{b,n})}. \quad (2.378)$$

By partial fraction expansion, there exist constants  $c_1, \dots, c_n \in \mathbb{R}$  for which,

$$\hat{G}(s) = \frac{c_1}{s - s_{b,1}} + \dots + \frac{c_n}{s - s_{b,n}}. \quad (2.379)$$

We recognize each term as the Laplace transform of an exponential signal,  $e^{s_{b,i}t}$ . Applying linearity of the Laplace transform, we conclude  $\hat{G}$  has an inverse Laplace transform.  $\square$

Thus, for the class of strictly proper, rational functions, we can always find an inverse Laplace transform. We'll provide an example of this shortly. First, we show how to apply the inverse Laplace transform to compute the zero-state response of an LTI system.

**Proposition 2.28 (Laplace Function Solution of I/O Systems)** Consider a continuous-time, LTI system with transfer function  $\hat{H} : \Omega \rightarrow \mathbb{C}^{p \times m}$ . The zero-state response of the system to an input signal  $u(\cdot) \in PC(\mathbb{R}_{\geq 0}, \mathbb{R}^m)$  of exponential order is computed,

$$y(t) = \mathcal{L}^{-1}(\hat{H}(s)U(s))(t), \quad (2.380)$$

for all  $t \in \mathbb{R}_{\geq 0}$  at which  $u$  is continuous, provided the inverse Laplace transform exists.

**Proof** We know that  $y(t) = (H * u)(t)$  for all  $t \geq 0$  at which  $u$  is continuous. Taking the Laplace transform, one then has  $Y(s) = \hat{H}(s)U(s)$ , which implies  $y(t) = \mathcal{L}^{-1}(\hat{H}(s)U(s))$ .  $\square$

Let's get some practice computing both the transfer function and the zero-state response of a continuous-time LTI system using the Laplace transform method.

*Example 2.7 (Step Response of a SISO System)* Consider the SISO, LTI system,

$$\dot{x}(t) = \begin{bmatrix} 0 & 1 \\ -1 & -2 \end{bmatrix} x(t) + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(t) \quad (2.381)$$

$$y(t) = [1 \ 0] x(t) \quad (2.382)$$

Let's compute the *step response* the system—the zero-state response of the system to a step function input. First, let's identify the transfer function of the system. There are a couple ways of doing this. First, we can use the formula  $\hat{H}(s) = C(sI - A)^{-1}B + D$ . This leaves us with,

$$\hat{H}(s) = [1 \ 0] \begin{bmatrix} s & -1 \\ 1 & s+2 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad (2.383)$$

which, since the system is small, we can compute using a symbolic calculator. Second, we can directly solve for the function  $\hat{H}$  which satisfies  $\hat{Y}(s) = \hat{H}(s)\hat{U}(s)$ . In order to do this, we examine the components of the state equation. We have,

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} x_2 \\ -x_1 - 2x_2 + u \end{bmatrix}. \quad (2.384)$$

Let's take the transform of each of component with  $x_i(0^-) = 0$ . We get,

$$s\hat{X}_1(s) = \hat{X}_2(s) \quad (2.385)$$

$$s\hat{X}_2(s) = -\hat{X}_1(s) - 2\hat{X}_2(s) + \hat{U}(s). \quad (2.386)$$

Computing the output, we find  $\hat{Y}(s) = \hat{X}_1(s)$ . Therefore, to get  $\hat{Y}(s) = \hat{H}(s)\hat{U}(s)$ , we solve,

$$s^2\hat{X}_1(s) = -\hat{X}_1(s) - 2s\hat{X}_1(s) + \hat{U}(s) \quad (2.387)$$

$$(s^2 + 2s + 1)\hat{X}_1(s) = \hat{U}(s) \quad (2.388)$$

$$\hat{X}_1(s) = \frac{1}{s^2 + 2s + 1}\hat{U}(s). \quad (2.389)$$

Plugging into the output, it follows that

$$\hat{Y}(s) = \frac{1}{s^2 + 2s + 1}\hat{U}(s) := \hat{H}(s)\hat{U}(s). \quad (2.390)$$

This formula for  $\hat{H}$  is identical to that which we would've found from the matrix inversion method. Note that the form of the state equation & output equations given here are particularly amenable to the direct solution method—later, we'll outline a general class of system representations for which this method is readily applicable.

Now that we've identified the transfer function of the system, we can compute its step response. We want to find the inverse Laplace transform of,

$$\hat{Y}(s) = \hat{H}(s)\hat{\mathbf{1}}(s) = \frac{1}{s^2 + 2s + 1} \frac{1}{s}. \quad (2.391)$$

A nice method to do this “by inspection” is to break up this transform into pieces composed of the common Laplace transforms we identified earlier. We'll accomplish this using partial fraction expansion<sup>12</sup>, which splits up the product of functions into a series of simpler fractions. With a little bit of algebra, we decompose  $\hat{H}(s)\hat{\mathbf{1}}(s)$  into,

<sup>12</sup> If you haven't seen this technique before, consult an introductory feedback control textbook such as [3] for the details on how to perform a partial fractions expansion by hand. Symbolic calculators such as Matlab symbolic have a function `partfrac` which will do this for you.

$$\frac{1}{s^2 + 2s + 1} \frac{1}{s} = \frac{1}{s} - \frac{1}{(s+1)^2} - \frac{1}{s+1}. \quad (2.392)$$

Now, we apply linearity of the Laplace transform and find the inverse transform of the entire expression by finding the inverse transform of each piece. Using the table of transforms and transform rules we derived earlier, we find,

$$\mathcal{L}^{-1}\left(\frac{1}{s}\right) = \mathbf{1}(t), \quad \mathcal{L}^{-1}\left(\frac{1}{(s+1)^2}\right) = te^{-t}, \quad \mathcal{L}^{-1}\left(\frac{1}{s+1}\right) = e^{-t}. \quad (2.393)$$

We conclude that the step response of the system is,

$$y(t) = \mathbf{1}(t) - te^{-t} - e^{-t}, \quad t \geq 0. \quad (2.394)$$

From start to finish, *all we needed* was a little bit of algebra—no fancy matrix exponentials or Jordan forms required!

In the example above, we found that the form of the  $(A, B, C, D)$  matrices allowed for a particularly simple solution of the transfer function. We now outline the general case of this simple structure.

**Proposition 2.29 (Representation with a Simple Transfer Function)** *Consider a SISO system representation,  $(A, B, C, D)$ , in which*

$$A = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & 1 \\ -a_0 & -a_1 & \dots & -a_{n-2} & -a_{n-1} \end{bmatrix} \quad B = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \quad (2.395)$$

$$C = [c_0 \ c_1 \ \dots \ c_{n-2} \ c_{n-1}] \quad D = 0.$$

*For such a system representation, the transfer function is computed,*

$$\hat{H}(s) = \frac{c_{n-1}s^{n-1} + c_{n-2}s^{n-2} + \dots + c_1s + c_0}{s^n + a_{n-1}s^{n-1} + \dots + a_1s + a_0}. \quad (2.396)$$

**Proof** See Problem 2.24. □

Later in the course, we'll discuss a more general version of Proposition 2.29 which lets us compute the transfer function of any SISO system representation using something called the *Markov parameters*. For now, however, this simple case will suffice.

Earlier, we mentioned that Proposition 2.28 enables one to compute the matrix exponential using an inverse Laplace transform. In the following result, we see exactly how to do this. What's more, we'll find that this method is *entirely independent* from the Jordan canonical form!

**Proposition 2.30 (Matrix Exponential via Inverse Laplace Transform)** *Consider a matrix  $A \in \mathbb{R}^{n \times n}$ . The exponential of  $At$  is computed via the inverse Laplace transform as*

$$\exp(At) = \left[ \mathcal{L}^{-1}[(sI - A)^{-1}] \right](t), \quad \forall t \geq 0. \quad (2.397)$$

**Proof** Consider the initial value problem,

$$\dot{X}(t) = AX(t), \quad X(0) = I, \quad (2.398)$$

the solution of which is the matrix exponential  $\exp(At)$ . Taking the Laplace transform,

$$s\hat{X}(s) - X(0) = A\hat{X}(s) \quad (2.399)$$

$$(sI - A)\hat{X}(s) = I \quad (2.400)$$

$$\hat{X}(s) = (sI - A)^{-1}. \quad (2.401)$$

Thus, the matrix exponential is the inverse Laplace transform of  $(sI - A)^{-1}$ .  $\square$

**Exercise 2.27** Compute the matrix exponential  $\exp(At)$  for the matrix

$$A = \begin{bmatrix} 0 & 1 \\ -1 & -2 \end{bmatrix} \quad (2.402)$$

using the inverse Laplace transform method.

Let's quickly summarize what we've found about Laplace transforms and compare it with our "Laplace transform wish list" that we came up with at the start of this section. So far, we've checked off the following two items:

1. Differential Equations Become Algebraic: we've shown that we can analyze the I/O response of state space equations using algebraic expressions that result from the Laplace transform. Additionally, we showed how to compute the matrix exponential using an entirely algebraic method.
2. Simplify Convolutions: we showed that, under the Laplace transform, convolution simply becomes *multiplication* of transforms. This is a significant simplification of the convolution operation.

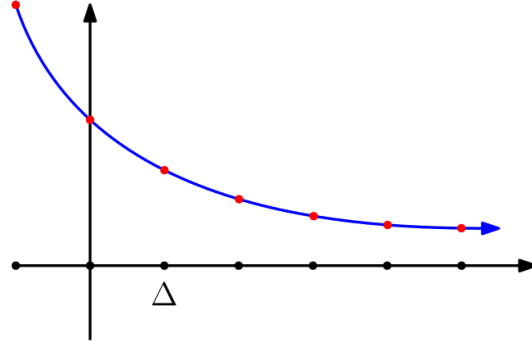
Currently, we have two items remaining on our wish list: *analyze long-term behavior* and *simple interpretations*. To fully answer these questions, we'll need to study stability theory (which we'll do in the next chapter) and realization theory (which we'll do in the chapter following that). In particular, we'll find that the transfer function encodes valuable information about the I/O stability of an LTI system, and that relationship between the numerator and denominator polynomials of a transfer function is wrapped up in the concepts of controllability and observability.

## 2.4.5 The $\mathcal{Z}$ -Transform

Now, we develop an analogous transform for discrete-time systems. The  $\mathcal{Z}$ -transform is the discrete-time analogue of the Laplace transform, and is defined to share many of its key properties with the Laplace transform. As such, much of the theory of the  $\mathcal{Z}$ -transform will immediately feel familiar to us.

The main insight into defining the  $\mathcal{Z}$ -transform is the following: the discrete-time analogue of an exponential signal,  $e^{at}$ , is the *geometric* signal,  $a^k$ . In fact, the geometric signal  $a^k$  simply comes from sampling an exponential signal  $e^{ct}$  at integer time steps,  $t = k\Delta$ . When

we combine the translation from exponential to geometric with the standard translation from integral to sum, we find a natural, candidate definition for the  $\mathcal{Z}$ -transform that mirrors that of the Laplace transform.



**Fig. 2.9** Sampling an exponential signal at a fixed sampling time step  $\Delta$  produces a geometric signal.

**Definition 2.44 (One-Sided  $\mathcal{Z}$ -Transform)** Consider a discrete-time, matrix-valued signal,  $f[\cdot] : \mathbb{Z}_{\geq 0} \rightarrow \mathbb{R}^{n \times m}$ . The one-sided  $\mathcal{Z}$ -transform of  $f[\cdot]$  is the complex-valued function,  $\hat{F} : \Omega \rightarrow \mathbb{C}^{n \times m}$ , defined

$$\hat{F}(z) = \mathcal{Z}[f](z) := \sum_{k=0}^{\infty} z^{-k} f[k], \quad (2.403)$$

where  $\Omega := \{z \in \mathbb{C} : \mathcal{Z}(f)(z) \text{ converges absolutely}\}$  is called the *region of absolute convergence* of the transform  $\hat{F}$ .

*Remark 2.64* Just like the one-sided Laplace transform, the one-sided  $\mathcal{Z}$ -transform starts at 0 and goes to  $\infty$  to reflect the causality of our systems (the discrete-time impulse response of a causal, LTI system is zero for  $k < 0$ ). One can, of course, extend the definition of the  $\mathcal{Z}$ -transform to  $-\infty$ . Here, we'll stick with the one-sided definition—whenever we say “ $\mathcal{Z}$ -transform,” we mean the one-sided  $\mathcal{Z}$ -transform.

*Remark 2.65* As with the case of the Laplace transform, we use a capital letter with a hat to distinguish between the  $\mathcal{Z}$ -transform of a signal and the original signal. Above, for instance, we write the  $\mathcal{Z}$ -transform of  $f[\cdot]$  as  $\hat{F}$ .

Take a moment to examine the formula for the  $\mathcal{Z}$ -transform, and convince yourself that *all we've done* is swap out the exponential signal  $e^{-st}$  for a geometric signal  $z^{-k}$  and an integral from  $0^-$  to  $\infty$  for a sum from 0 to  $\infty$ . In discrete time, of course,  $0^-$  is no different from 0 (since we're working in  $\mathbb{Z}$  and not  $\mathbb{R}$ ), so our sum starts from 0.

Now that we've established a definition for the  $\mathcal{Z}$ -transform, let's try tracing a few more steps we took when defining the Laplace transform. Recall that, after defining the Laplace transform, we sought out a class of signals with a well-defined Laplace transform, which turned out to be the signals of *exponential order*—this followed from the presence of an exponential in the definition of the transform. In the definition of the  $\mathcal{Z}$ -transform, instead of an exponential, we have a geometric growth term,  $z^{-k}$ . The following definition is therefore a natural candidate for a class of signals with well-defined  $\mathcal{Z}$ -transforms.

**Definition 2.45 (Sequence of Geometric Order)** A sequence  $f : \mathbb{Z}_{\geq 0} \rightarrow \mathbb{R}^{p \times m}$  is of geometric order  $\alpha \geq 0$  if there exists an  $M \geq 0$  such that, for all  $k \in \mathbb{Z}_{\geq 0}$ ,  $\|f[k]\| \leq M\alpha^k$ .

*Remark 2.66* Unlike in the continuous-time case, we request that the order  $\alpha$  be nonnegative. If we don't make this restriction,  $\alpha^k$  could be negative for odd  $k$ .

**Exercise 2.28** Confirm that the choice of norm  $\|\cdot\|$  does not affect the order  $\alpha$  of a given sequence. Conclude that Definition 2.45 does not depend on the choice of norm.

With this definition in hand, we now state and prove a result concerning the convergence of the  $\mathcal{Z}$ -transform of sequences of geometric order.

**Proposition 2.31 ( $\mathcal{Z}$ -Transform of a Sequence of Geometric Order)** Consider a sequence  $f[\cdot] : \mathbb{Z}_{\geq 0} \rightarrow \mathbb{R}^{p \times m}$  of geometric order  $\alpha \geq 0$ . The region of absolute convergence of the  $\mathcal{Z}$ -transform of  $f[\cdot]$  contains the set

$$\{z \in \mathbb{C} : |z| > \alpha\} \subseteq \mathbb{C}. \quad (2.404)$$

**Proof** Suppose  $f$  is a sequence of geometric order  $\alpha \geq 0$ . Then, one has

$$\sum_{k=0}^{\infty} |z^{-k}| \|f[k]\| \leq M \sum_{k=0}^{\infty} |z^{-k}| \alpha^k = M \sum_{k=0}^{\infty} |z^{-1}\alpha|^k. \quad (2.405)$$

If  $|z^{-1}\alpha| < 1$ , this series will converge. This occurs for  $|z^{-1}| \cdot |\alpha| < 1 \iff |z| > \alpha$ .  $\square$

As opposed to the Laplace transform, where regions of convergence of functions of exponential order were of the form  $\text{Re}(s) > \alpha$ , for  $\mathcal{Z}$ -transforms and sequences of geometric order, one has  $|z| > \alpha$ —that the  $\mathcal{Z}$ -transform converges outside a disk centered at the origin. Now that we've outlined this “nice” class of signals, we consider some common  $\mathcal{Z}$ -transforms.

**Theorem 2.21 (Common One-Sided  $\mathcal{Z}$ -Transforms)** Consider the following collection of signals, transforms, and regions of absolute convergence of their transforms.

| Signal Name  | Signal  | $\mathcal{Z}$ -Transform   | R.O.C.       |
|--------------|---|--|--------------|
| Unit Impulse | $\delta[k]$   | 1  | $\mathbb{C}$ |
| Unit Step    | $\mathbf{1}[k] = \begin{cases} 1 & k \geq 0 \\ 0 & k < 0 \end{cases}$ | $\frac{1}{1-z^{-1}}$   | $ z  > 1$    |
| Geometric    | $a^k$   | $\frac{1}{1-az^{-1}}$  | $ z  >  a $  |
|              | $\sin(\omega k)$  | $\frac{\sin(\omega)z^{-1}}{1-2\cos(\omega)z^{-1}+z^{-2}}$        | $ z  > 1$    |
|              | $\cos(\omega k)$  | $\frac{1-\cos(\omega)z^{-1}}{1-2\cos(\omega)z^{-1}+z^{-2}}$      | $ z  > 1$    |
|              | $a^k \sin(\omega k)$  | $\frac{1-a\sin(\omega)z^{-1}}{1-2a\cos(\omega)z^{-1}+a^2z^{-2}}$ | $ z  >  a $  |
|              | $a^k \cos(\omega k)$  | $\frac{1-a\cos(\omega)z^{-1}}{1-2a\cos(\omega)z^{-1}+a^2z^{-2}}$ | $ z  >  a $  |



*Remark 2.67* Note that here, we assume all signals are zero for  $k < 0$ . One can multiply each signal by the unit step  $\mathbb{1}[k]$  to make this more explicit.

*Remark 2.68* Even though we've switched to discrete-time, the same remarks about analytic continuation still apply, as the transforms are still nothing more than complex functions. We can use analytic continuation to unambiguously extend the domain of each transform beyond the region of convergence of its corresponding infinite sum.

Thus, we observe that the set of common, one-sided  $\mathcal{Z}$ -transforms have regions of absolute convergence that are *punctured planes*—complex planes with a disk removed. As with the Laplace transform, we work out a couple of important examples of  $\mathcal{Z}$ -transforms.

*Example 2.8 ( $\mathcal{Z}$ -Transform of a Unit Impulse)* Since  $\delta[k]$  is nonzero only for  $k = 0$ ,

$$\mathcal{Z}(\delta)(z) = \sum_{k=0}^{\infty} z^{-k} \delta[k] = z^0 = 1. \quad (2.406)$$

It follows that  $(\mathcal{Z}(\delta))(z) = 1$  for all  $z \in \mathbb{C}$ , and that the transform converges for all  $z \in \mathbb{C}$ .

*Example 2.9 ( $\mathcal{Z}$ -Transform of a Unit Step)* Now, we consider the unit step function, which is identically 1 for  $k \geq 0$ . We have,

$$\sum_{k=0}^{\infty} z^{-k} \mathbb{1}[k] = \sum_{k=0}^{\infty} z^{-k}, \quad (2.407)$$

which converges absolutely for  $|z| > 1$ . Recalling the formula for an infinite geometric series, we conclude that for  $|z| > 1$ ,

$$(\mathcal{Z}(\mathbb{1}))(z) = \sum_{k=0}^{\infty} (z^{-1})^k = \frac{1}{1 - z^{-1}}. \quad (2.408)$$

**Exercise 2.29** Complete the  $\mathcal{Z}$ -transform table in Theorem 2.21. See [21] (or any other standard book on signals and systems) for a solution.

Now, we state a theorem regarding the most important properties of the  $\mathcal{Z}$ -transform. Since the  $\mathcal{Z}$ -transform was constructed in a similar spirit to the Laplace transform, all of the same key properties hold! Due to similarities to the analogous Laplace transform results, we leave the proof of the following theorem as an exercise.

**Theorem 2.22 (Key Properties of the  $\mathcal{Z}$ -Transform)** Let  $f, g : \mathbb{Z}_{\geq 0} \rightarrow \mathbb{R}$  be sequences of geometric orders  $\alpha$  and  $\beta$ , respectively.

1. *Analyticity:* at any  $|z| > \alpha + \epsilon$ ,  $\epsilon > 0$ ,  $\mathcal{Z}(f)(z)$  is analytic.
2. *Linearity:*  $\mathcal{Z}(k_1 f + k_2 g)(z) = k_1 \mathcal{Z}(f)(z) + k_2 \mathcal{Z}(g)(z) \quad \forall k_1, k_2 \in \mathbb{R}, z : |z| > \max\{\alpha, \beta\}$ .
3. *Convolution:*  $\mathcal{Z}(f * g)(z) = \mathcal{Z}(f)(z) \cdot \mathcal{Z}(g)(z)$  for all  $z$  satisfying  $|z| > \max\{\alpha, \beta\}$ .

**Exercise 2.30** Prove Theorem 2.22.

Thus, we observe that—as with the Laplace transform—the  $\mathcal{Z}$ -transform satisfies a number of desirable properties. First, the  $\mathcal{Z}$ -transform is linear—this means that we can easily take  $\mathcal{Z}$ -transforms of linear combinations of a few simple signals. Secondly, we observe that

convolution becomes *multiplication* under the  $\mathcal{Z}$ -transform. Recall that—in the case of the Laplace transform—this is the key property that enabled the use of transfer functions to study LTI systems. Since we have the same convolution property in the case of the  $\mathcal{Z}$ -transform, we'll find that the same definitions and properties of transfer functions hold for the  $\mathcal{Z}$ -transform. Before we turn our attention to this, however, we state some further important properties of the  $\mathcal{Z}$ -transform.

**Theorem 2.23 (Further Properties of the  $\mathcal{Z}$ -Transform)** *Let  $f : \mathbb{Z}_{\geq 0} \rightarrow \mathbb{R}$  be a signal of geometric order  $\alpha$  with  $\mathcal{Z}$ -transform  $F(z)$  on  $\Omega$ . Then, the following properties hold on a nonempty subset of  $\Omega$ .*

| Operation               | Signal     | $\mathcal{Z}$ -Transform                | (2.409) |
|-------------------------|------------|---|---------|
| Step Forward by 1       | $f[k+1]$   | $zF(z) - zf[0]$                         |         |
| Step Forward by $n > 0$ | $f[k+n]$   | $z^n F(z) - z^n f[0] - \dots - zf[k-1]$ |         |
| Delay by $n > 0$        | $f[k-n]$   | $z^{-n} F(z)$                           |         |
| Time Reversal           | $f[-k]$    | $F(z^{-1})$                             |         |
| Product with $k$        | $kf[k]$    | $-kF'(z)$                               |         |
| Scaling by $a^k$        | $a^k f[k]$ | $F(a^{-1}z)$                            |         |

**Exercise 2.31** Prove Theorem 2.23. See [21] (or any other standard book on signals and systems) for a solution.

### 2.4.5.1 Transfer Functions

We'll finish up our first pass at  $\mathcal{Z}$ -transforms by discussing the transfer functions of discrete-time, LTI systems. Due to the convolution property of the  $\mathcal{Z}$ -transform, we once again find that knowing the  $\mathcal{Z}$ -transform of the impulse response of an LTI system is sufficient to characterize its zero-state response to any input. As such, we define the *transfer function*, which entirely characterizes the system's response, as follows.

**Definition 2.46 (Discrete-Time Transfer Function)** Consider a discrete-time, LTI system with LTI impulse response map  $H[\cdot] : \mathbb{Z} \rightarrow \mathbb{R}^{p \times m}$ . The transfer function of the system is the map  $\hat{H} : \Omega \subseteq \mathbb{C} \rightarrow \mathbb{C}^{p \times m}$ ,

$$\hat{H}(z) = \mathcal{Z}(H)(z), \quad \forall z \in \Omega, \quad (2.410)$$

where  $\Omega$  is the region of convergence of the transform.

As in the continuous-time case, the discrete-time transfer function is equivalently characterized by its interaction with input signals and the zero-state response.

**Lemma 2.16 (Characterizing the Transfer Function)** *Consider a discrete-time, LTI system representation  $(A, B, C, D)$  with LTI impulse response map  $H[\cdot]$ . A function  $\hat{H} : \Omega \rightarrow \mathbb{C}^{p \times m}$  is the transfer function of the system if and only if  $\hat{Y}(z) = \hat{H}(z)\hat{U}(z)$  for all transforms  $\hat{U}(z)$  and  $\hat{Y}(z)$  of admissible input signals  $u[\cdot]$  and their corresponding zero-state responses  $y[\cdot]$ .*

**Proof** The proof of this result follows identically to the continuous-time case. If  $\hat{H}$  is the transfer function of the system, then the convolution property of the  $\mathcal{Z}$ -transform necessitates  $\hat{Y}(z) = \hat{H}(z)\hat{U}(z)$  for all pairs  $u[\cdot], y[\cdot]$  of input and zero-state response. Now, suppose  $\hat{H}$  satisfies the given property. Consider an input signal  $u[\cdot] = \delta[\cdot]e_j$ . Then,  $\hat{Y}(z) = \hat{H}(z)e_j$ . Since  $\hat{Y}(z)$  must also equal the transform of column  $j$  of the impulse response map, it follows that  $\hat{H}(z)$  must be the transform of the impulse response map.  $\square$

As in the continuous-time case, this result makes the computation of the transfer function of an LTI system quite simple.

**Proposition 2.32 (Transfer Function of a DT-LTI System Representation)** *Consider a discrete-time LTI system representation  $(A, B, C, D)$ . The transfer function of the system is computed,*

$$\hat{H}(z) = C(zI - A)^{-1}B + D, \quad \forall z \in \Omega, \quad (2.411)$$

where  $\Omega = \mathbb{C} \setminus \text{spec}(A)$ , the complex plane minus the eigenvalues of  $A$ .

**Remark 2.69** When proving statements about the  $\mathcal{Z}$ -transform, it's useful to draw comparisons to analogous theorems about the Laplace transform. In the case above, for instance, the “step ahead by 1” operation in discrete-time is the analogue of the derivative in continuous-time. This is reflected in their transforms—step ahead by 1 corresponds to multiplication by  $z$ , whereas differentiation corresponds to multiplication by  $s$ . Observations like this often give away the entire proof structure of a property of the  $\mathcal{Z}$ -transform once the analogous Laplace transform proof has been completed.

**Exercise 2.32** Prove Proposition 2.32, using the continuous-time case as a guide.

Once we've defined the transfer function of the system, we can determine the system response in the *time-domain* to an input signal using the inverse  $\mathcal{Z}$ -transform.

**Definition 2.47 (Inverse  $\mathcal{Z}$ -Transform)** The inverse  $\mathcal{Z}$ -transform is a map  $\mathcal{Z}^{-1}$  taking a complex-valued function  $\hat{H} : \Omega \rightarrow \mathbb{C}^{p \times m}$  to a real-valued sequence  $H[\cdot] : \mathbb{Z}_{\geq 0} \rightarrow \mathbb{C}^{p \times m}$  satisfying  $\mathcal{Z}(H)(z) = \hat{H}(z)$  for all  $z \in \Omega$ .

**Lemma 2.17 (Inverse  $\mathcal{Z}$ -Transform of a Strictly Proper Rational Function)** *Consider a complex function which is the ratio of two polynomials with real coefficients,*

$$\hat{G}(z) = \frac{a_m z^m + \dots + a_1 z + a_0}{b_n z^n + \dots + b_1 z + b_0}, \quad a_i, b_i \in \mathbb{R}, \quad a_m, b_n \neq 0. \quad (2.412)$$

*If  $m < n$ , then  $\hat{G}$  has a unique inverse  $\mathcal{Z}$ -transform.*

**Exercise 2.33** Provide a proof of Lemma 2.17 in the special case where the numerator and denominator polynomials each have distinct, real roots.

We can compute the response of a system to any input using knowledge of the system's transfer function and the input's  $\mathcal{Z}$ -transform.

**Proposition 2.33 ( $\mathcal{Z}$ -Transform Solution of I/O Systems)** *Consider a discrete-time, LTI system with transfer function  $\hat{H} : \Omega \rightarrow \mathbb{C}^{p \times m}$ . The zero-state response of the system to an input signal  $u[\cdot] : \mathbb{Z}_{\geq 0} \rightarrow \mathbb{R}^m$  of geometric order is computed,*

$$y[k] = \mathcal{Z}^{-1}(\hat{H}(z)\hat{U}(z))[k], \quad \forall k \in \mathbb{Z}_{\geq 0}, \quad (2.413)$$

*provided the inverse  $\mathcal{Z}$ -transform exists.*

*Remark 2.70* Since continuity of the input signal is not an issue in discrete-time, we no longer need to append the condition “at every  $t$  at which  $u$  is continuous.”

**Exercise 2.34** Prove Proposition 2.33.

Just as we can compute a matrix exponential using the inverse Laplace transform, we can compute a matrix *power* using the inverse  $\mathcal{Z}$ -transform.

**Proposition 2.34 (Matrix Power via Inverse  $\mathcal{Z}$ -Transform)** *Consider a matrix  $A \in \mathbb{R}^{n \times n}$ . The exponent  $A^k$ ,  $k \in \mathbb{Z}_{\geq 0}$  is computed via the inverse  $\mathcal{Z}$ -transform as*

$$A^k = \left[ \mathcal{Z}^{-1}[z(zI - A)^{-1}] \right][k], \quad \forall k \in \mathbb{Z}_{\geq 0}. \quad (2.414)$$

**Exercise 2.35** Prove Proposition 2.34. Remember to account for the initial condition!

Let's summarize what we've found about the  $\mathcal{Z}$ -transform. Across the board, we've found that the  $\mathcal{Z}$ -transform mirrors the Laplace transform in discrete-time. We found that the  $\mathcal{Z}$ -transform is linear, turns convolution into multiplication, and interacts well with standard operations such as time shifts and delays. Like the Laplace transform, we'll return to a deeper study of the  $\mathcal{Z}$ -transform later in the course.

## 2.4.6 Further Reading

The treatment of impulse response in discrete and continuous-time is based on [2] and [8]. The informal approach to the Dirac delta “function” as a function of time is based on the treatment given by [2]. For a more formal treatment of the Dirac delta distribution, we refer the reader to [22] for a nice, self-contained exposition. For more information on approximations to the identity, we recommend [37] and [39]. A nice book on complex analysis is [38]. For further information on the analytical properties of the Laplace and  $\mathcal{Z}$ -transforms, as well as on a rigorous construction of distributions within the context of control, the reader is encouraged to consult [22, 23, 24]. A more engineering-oriented introduction to the Laplace and  $\mathcal{Z}$ -transforms is provided in [21], [28], and [41].

### 2.4.7 Problems

**Problem 2.21 (Algebraic Properties of Convolutions)** In this problem, we'll examine the basic algebraic properties of convolutions. For both the continuous and discrete-time convolutions, confirm that the following properties hold:

1. Linearity:  $(\alpha f_1 + \beta f_2) * g = \alpha f_1 * g + \beta f_2 * g$ .
2. Commutativity:  $f * g = g * f$ .
3. Associativity:  $f * (g * h) = (f * g) * h$ .

You may assume that each convolution (e.g.  $f * g$ ,  $g * h$ , etc.) is defined for all time.

**Problem 2.22 (Transforms & Transition Matrices)** The Laplace transform offers yet another way of computing the state transition matrix. For a continuous-time, LTI representation  $(A, B, C, D)$ , the matrix exponential is computed  $\exp(At) = \mathcal{L}^{-1}[(sI - A)^{-1}](t)$ , for all  $t \geq 0$ . In this problem, we'll consider an analogue in discrete-time, and use both the continuous and discrete formulas to compute some transition matrices.

1. Show that the state transition matrix of a discrete-time, LTI representation  $(A, B, C, D)$  is computed,

$$A^k = \mathcal{Z}^{-1}[z(zI - A)^{-1}][k], \quad \forall k \geq 0. \quad (2.415)$$

2. Using the transform formulas, compute the continuous and discrete-time transition matrices associated to the matrix,

$$A = \begin{bmatrix} 0 & 1 \\ -1 & -2 \end{bmatrix}. \quad (2.416)$$

Comment on the benefits and drawbacks of this method of computing the transition matrix. You may use a symbolic calculator to compute the inverse of  $(sI - A)$ .

**Problem 2.23 (Transfer Functions & Change of Basis)** Consider a linear, time-invariant system representation  $(A, B, C, D)$ . Recall that under a change of state coordinates,  $z = Tx$ , the representation *transforms* to  $(TAT^{-1}, TB, CT^{-1}, D)$ . Does the transfer function associated to the system representation change under a change of state coordinates? Provide a proof or counterexample to back up your answer.

**Problem 2.24 (A Simple SISO Transfer Function)** Consider a continuous-time SISO, LTI system representation  $(A, B, C, D)$ ,

$$A = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & 1 \\ -a_0 & -a_1 & \dots & -a_{n-2} & -a_{n-1} \end{bmatrix} \quad B = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \quad (2.417)$$

$$C = [c_0 \ c_1 \ \dots \ c_{n-2} \ c_{n-1}] \quad D = 0,$$

Show that the transfer function of such a system is computed,

$$\hat{H}(s) = \frac{c_{n-1}s^{n-1} + c_{n-2}s^{n-2} + \dots + c_1s + c_0}{s^n + a_{n-1}s^{n-1} + \dots + a_1s + a_0}. \quad (2.418)$$

**Problem 2.25 (Some Laplace &  $\mathcal{Z}$ -Transforms)** In this problem, we'll establish a couple of basic Laplace and  $\mathcal{Z}$ -transforms.

1. Let  $f$  be a signal and  $\tau > 0$ . Define the signal  $g$ ,

$$g(t) = \begin{cases} 0, & 0 \leq t < \tau. \\ f(t - \tau), & t \geq \tau. \end{cases} \quad (2.419)$$

Show that  $\hat{G}(s) = e^{-s\tau} \hat{F}(s)$ .

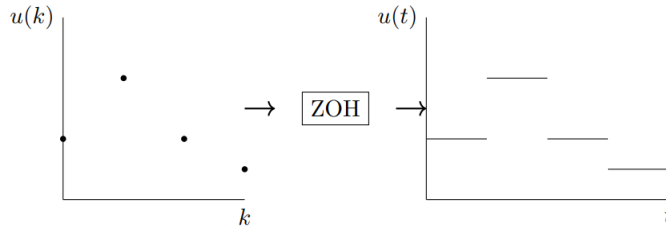
2. Show that the  $\mathcal{Z}$ -transform of the unit step function,

$$\mathbb{1}[k] = \begin{cases} 1, & k \geq 0 \\ 0, & k < 0, \end{cases} \quad (2.420)$$

is  $\hat{\mathbb{1}}(z) = z/(z - 1)$ .

**Problem 2.26 (Sampled-Data Systems [40])** In this problem, we'll take a frequency-domain approach to zero-order hold discretization. Consider a continuous-time system where inputs and outputs can only be accessed at discrete times  $t = k\Delta$ ,  $k \in \mathbb{Z}$ , with a sampling period  $\Delta \in \mathbb{R}_{>0}$ . The discrete-time input  $u[k] = u(k\Delta)$  is passed through a zero-order hold (ZOH) digital-to-analog (D/A) converter that accepts the input  $u(k\Delta)$  at  $t = k\Delta$  and holds it constant until the next input is applied at  $t = (k + 1)\Delta$ . The continuous-time system processes the ZOH output, and its resulting continuous-time output is sampled by the A/D converter to produce the discrete-time output  $y[k]$ . The goal is to compute the discrete-time transfer function of the overall system considering the effects of D/A and A/D converters.

1. First, we will compute the transfer function of the zero order hold. The zero-order hold takes in a continuous-time input signal (which has been sampled with interval  $\Delta$ ) and returns a continuous-time *held* version of the signal.

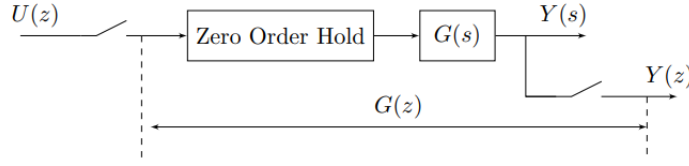


Show that the transfer function of the zero-order hold is

$$H(s) = \frac{1 - e^{-s\Delta}}{s}. \quad (2.421)$$

Assume that the sampler translates a discrete-time unit impulse to a continuous-time Dirac delta.

2. Now, we apply the zero-order hold block to a system. Consider the block diagram,



where  $G(s)$  is the transfer function of a continuous-time system and the latches represent sampling. Show that the transfer function from  $U(z)$  to  $Y(z)$  is,

$$G(z) = (1 - z^{-1}) \mathcal{Z} \left[ \mathcal{L}^{-1} \left( \frac{G(s)}{s} \right) \right]. \quad (2.422)$$

*Hint: Split  $H(s)G(s)$  into two components. How do the two components relate?*

3. If the continuous-time transfer function is,

$$G(s) = \frac{a}{s^2}, \quad (2.423)$$

what is the corresponding discrete-time transfer function  $G(z)$ ?

**Problem 2.27 (The  $\mathcal{Z}$ -Transform & the Fibonacci Sequence)** The Fibonacci sequence is defined recursively as,

$$x[0] = 0, x[1] = 1, x[k+1] = x[k] + x[k-1], \forall k \geq 1. \quad (2.424)$$

Compute the  $\mathcal{Z}$ -Transform of the Fibonacci sequence. Then, calculate the roots of the denominator of the  $\mathcal{Z}$ -transform. Do you recognize any of the roots? *Look up the “golden ratio” if you do not.*

**Problem 2.28 (Analytic Functions)** Recall that a given function  $f : \Omega \rightarrow \mathbb{C}$ , where  $\Omega \subseteq \mathbb{C}$  is open in  $\mathbb{C}$ , is an *analytic function* if it is (complex) differentiable in a neighborhood of every point of  $\mathbb{C}$ . For each of the *scalar* functions,

$$f_1(s) = \frac{1}{s}, f_2(s) = e^s, f_3(s) = \frac{(s-1)}{(s+1)(s-1)(s+2)}, G(s) = C(sI - A)^{-1}B, \quad (2.425)$$

determine the largest subset of  $\mathbb{C}$  on which the function is analytic.





## Chapter 3

# Stability of Linear Systems

In the previous chapter, we studied linear dynamical systems, their representations, and the structure of their solutions. In this chapter, we move on to the study of *stability*. What does it mean for a linear dynamical system to be *stable*?

We'll take two approaches in answering this question—the state space and the I/O approaches. In the first half of this chapter, we'll study stability from the state space perspective. In particular, we will study the stability of the unforced systems  $\dot{x} = A(t)x$ ,  $x[k+1] = A[k]x[k]$ , by examining properties of their state transition maps. Following this, we'll reintroduce inputs, and study input/output stability. Here, we'll answer the question: given a bounded *input* to a system, when will the *output* be bounded? To give a satisfying answer to this question, we'll develop some more transfer function theory, and will discuss norms on input/output systems.

### 3.1 The Analytical Approach to Stability

In this section, we begin our study of stability of linear systems from the *state space* perspective. First, we'll formulate a definition for stability for such systems, using the concept of an equilibrium point. Following this, we'll work towards an exact characterization of stability for linear, time-varying systems, using the state transition matrix. With this characterization in hand, we'll proceed to show that the stability of a linear, time-invariant system is *entirely* determined by the eigenvalues of the representation matrix  $A$ .

As the title of this section suggests, the approach we'll take in proving the results in this section is fundamentally *analytical*. Instead of taking a more algebraic approach, we'll determine the stability of linear systems by formulating bounds directly on the norm of the state,  $\|x(t)\|$ . As we move through this chapter, we'll discover that the direct analytical approach we take in this section is but one of many ways of studying stability.

#### 3.1.1 Defining Stability

As a first step towards characterizing the stability of linear systems, we must find a precise, mathematical answer to the question,

*What does it mean for a system to be stable?*

Despite the innocent nature of this question, finding a satisfying answer takes some work. As it happens, there are a number of formal definitions that answer this question. One can pick from a range of definitions of stability—input/output stability, Lyapunov stability, structural stability, Poisson stability—the list goes on. Dynamicists have been rather busy!

In this section, we'll focus on *Lyapunov stability*, perhaps the most relevant definition of stability in the state space picture of control. Let's define the setting for our study of Lyapunov stability. In this section, we'll remove inputs from the picture entirely, and study the Lyapunov stability of the *unforced* linear systems,

$$\dot{x}(t) = A(t)x(t) \text{ and } x[k+1] = A[k]x[k], \quad (3.1)$$

linear systems in which the evolution of the state does not depend on an input signal  $u(\cdot)$ . As we gain a more sophisticated understanding of stability, we'll come to appreciate that the unforced case reveals a great deal about the forced case. The interplay between stability of systems with and without inputs will become a central theme of our study.

Since we'll be working entirely *without* inputs in this section, we'll introduce some new notation to relieve ourselves of some of the overhead that comes with working with inputs. Since the evolution of an unforced system does not depend on an input signal, we define the *unforced state transition map*  $\varphi : \mathbf{T} \times \Sigma \rightarrow \Sigma$ ,  $\mathbf{T} = \{(t_1, t_0) \in \mathcal{T} \times \mathcal{T} : t_1 \geq t_0\}$ , as

$$\varphi(t, t_0, x_0) = \varphi(t, t_0, x_0, 0). \quad (3.2)$$

This way, we don't have to carry around an (irrelevant) input signal. Based on this notation, we have that, for a system with state equation  $\dot{x}(t) = A(t)x(t)$ ,  $\varphi(t, t_0, x_0)$  represents the unique solution to the state equation at time  $t$ , starting from  $x(t_0) = x_0$ . Likewise, for  $x[k+1] = A[k]x[k]$ ,  $\varphi(k, k_0, x_0)$  represents the solution to the state equation at time  $k$ , starting from  $x[k_0] = x_0$ .

Now that we've introduced this simplifying notation, we turn back to the study of stability. How might we formulate a definition of stability for an unforced, state space system? First, it seems logical that a system can have different stability properties at different points in its state space. Near one point in the state space, a system might act in a very stable and predictable manner, while near another point, it might exhibit a wholly different behavior. As a consequence of this, we are not interested in the stability of an entire "system," but rather in the stability of special, distinguished points in the state space of the system. In particular, we are interested in the stability of *equilibrium points*.

**Definition 3.1 (Equilibrium Point)** Consider an unforced system with state transition map  $\varphi : \mathbf{T} \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ . A point  $x_e \in \mathbb{R}^n$  is an *equilibrium point* of the system if,

$$x_e = \varphi(t, t_0, x_e), \quad \forall t \geq t_0 \in \mathcal{T}. \quad (3.3)$$

*Remark 3.1* Notice how this definition holds for *both* discrete and continuous-time systems, since we use a general time set  $\mathcal{T}$ !

Thus, we observe that a point is said to be an *equilibrium point* of an unforced system if it is a *fixed point* of the state transition map—once we get to an equilibrium point, we stay there for all time. This fixed point property illustrates why equilibrium points naturally arise in the study of stability. Since we will stray away from any non-equilibrium point

as time passes—a characteristic which is intuitively unstable—we rule out non-equilibrium points from our study of stability. Let’s use the state equations to come up with equivalent characterizations of equilibria in the continuous and discrete-time cases.

**Proposition 3.1 (Characterization of Equilibria)** *The equilibria of continuous and discrete-time unforced systems are characterized as follows.*

1. *Continuous-time:* a point  $x_e \in \mathbb{R}^n$  is an equilibrium point of the unforced, CT-LTV system  $\dot{x}(t) = A(t)x(t)$  if and only if  $A(t)x_e = 0$ , for all  $t \in \mathbb{R}$  at which  $A(\cdot)$  is continuous.
2. *Discrete-time:* a point  $x_e \in \mathbb{R}^n$  is an equilibrium point of the unforced, DT-LTV system  $x[k+1] = A[k]x[k]$  if and only if  $x_e = A[k]x_e$ , for all  $k \in \mathbb{Z}$ .

**Proof** First, we’ll prove the continuous-time case. Suppose  $x_e$  is an equilibrium point in the sense of Definition 3.1. Fix a time  $t \in \mathbb{R}$  and a time  $t_0 < t$ . Then, for any  $\epsilon > 0$ , one has,

$$x_e = \varphi(t + \epsilon, t_0, x_e). \quad (3.4)$$

This implies that,

$$\frac{\varphi(t + \epsilon, t_0, x_e) - \varphi(t, t_0, x_e)}{\epsilon} = \frac{x_e - x_e}{\epsilon} = 0. \quad (3.5)$$

Taking the limit of both sides as  $\epsilon \rightarrow 0$ , we conclude,

$$\frac{d}{dt}\varphi(t, t_0, x_e) = 0, \quad (3.6)$$

Thus, it must be that  $0 = \dot{\varphi}(t, t_0, x_e) = A(t)\varphi(t, t_0, x_e) = A(t)x_e$  at all  $t$  at which  $A(\cdot)$  is continuous. This shows the first direction. Now, for the other direction, suppose  $A(t)x_e = 0$  for all  $t \in \mathbb{R}$ . Then, for any  $t \geq t_0$ , one has  $\varphi(t, t_0, x_e) = x_e$ , since the derivative of the state transition map is zero on any compact interval for all but a finite number of points.

Next, we show the discrete-time case. First, suppose  $x_e = \varphi(k, k_0, x_e)$  for all  $k \geq k_0$ . Fix  $k \in \mathbb{Z}$ . Then, it follows that  $\varphi(k+1, k, x_e) = A[k]x_e = x_e$ , which completes the first direction. The opposite direction proceeds by induction on  $k$ .  $\square$

**Exercise 3.1** Complete the second direction of item (2) by induction on  $k$ .

As a consequence of Proposition 3.1, we can equivalently define the equilibria of continuous-time, LTV systems as points satisfying  $A(t)x_e = 0$ , and those of discrete-time systems as points satisfying  $A[k]x_e = x_e$ . These results lend the following, obvious candidate for an equilibrium point of a linear system.

**Proposition 3.2 (Zero is Always an Equilibrium)** *The point  $x_e = 0 \in \mathbb{R}^n$  is an equilibrium point of an unforced linear system.*

**Exercise 3.2** Verify the claim of Proposition 3.2.

Now that we’ve defined what equilibria actually are, we devote our attention to defining their stability. Intuitively, what should it mean for an equilibrium point of a system to be stable? We capture the basic idea of stability with the following: an equilibrium point is *stable* if starting close to the equilibrium means we will *stay close* to the equilibrium for all time. This intuitive “start close, stay close” notion of stability, as well as some of its important variants, are made precise in the following definition.

**Definition 3.2 (Lyapunov/Asymptotic Stability)** Consider an unforced system with state transition map  $\varphi : \mathbf{T} \times \mathbb{R}^n \rightarrow \mathbb{R}^n$  and equilibrium point  $x_e \in \mathbb{R}^n$ . The point  $x_e$  is:

1. Lyapunov stable (SISL): if for all  $t_0 \in \mathcal{T}$  and  $\epsilon > 0$ ,  $\exists \delta > 0$  such that,

$$\|x_0 - x_e\| < \delta \implies \|\varphi(t, t_0, x_0) - x_e\| < \epsilon, \forall t \geq t_0. \quad (3.7)$$

2. Locally asymptotically stable (LAS): if it is SISL and for all  $t_0 \in \mathcal{T}$ ,  $\exists \epsilon_0 > 0$  such that,

$$\|x_0 - x_e\| < \epsilon_0 \implies \lim_{t \rightarrow \infty} \varphi(t, t_0, x_0) = x_e. \quad (3.8)$$

3. Globally asymptotically stable (GAS): if it is SISL and for all  $t_0 \in \mathcal{T}$ ,  $x_0 \in \mathbb{R}^n$ ,

$$\lim_{t \rightarrow \infty} \varphi(t, t_0, x_0) = x_e. \quad (3.9)$$

4. Unstable: if it is not Lyapunov stable.<sup>1</sup>

*Remark 3.2* Lyapunov stability is sometimes referred to as *stability in the sense of Lyapunov*, which lends Lyapunov stability the abbreviation *SISL*. Since Lyapunov stability is the most “basic” definition of stability of an equilibrium point, it is also often referred to simply as *stability*.

*Remark 3.3* We’re generally only interested in one type of instability—that which contradicts the definition of Lyapunov stability. Since asymptotic stability requires Lyapunov stability, a point which is unstable is automatically not asymptotically stable.

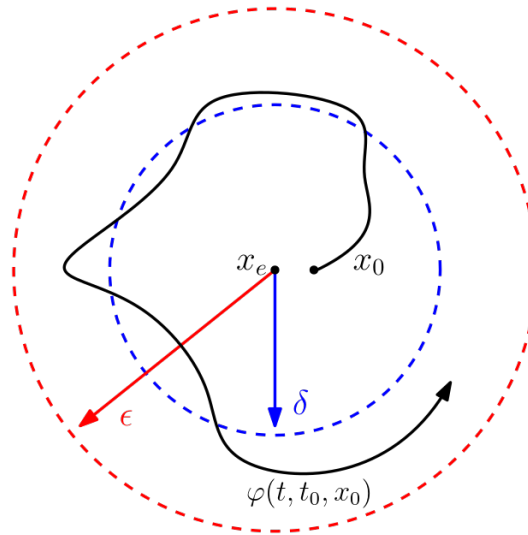
**Exercise 3.3** Verify that each of the definitions of stability *does not* depend on the choice of norm,  $\|\cdot\|$ , on  $\mathbb{R}^n$ .

Let’s run through the different definitions of stability we presented above. The first definition, *Lyapunov stability*, exactly captures our “start close, stay close” idea of stability. It says, for *any* initial time  $t_0 \in \mathbb{R}$  and specification  $\epsilon > 0$  of how close we’d like to remain to our equilibrium, we can always find a distance  $\delta > 0$  (possibly dependent on *both*  $t_0$  and  $\epsilon$ ) such that if we start a distance  $\delta$  away from our equilibrium, we’ll remain within a distance  $\epsilon$  for all time. In other words, if we *start close* (within a distance  $\delta$ ), we’ll *stay close* (within a distance  $\epsilon$ ) for all  $t \geq t_0$ .

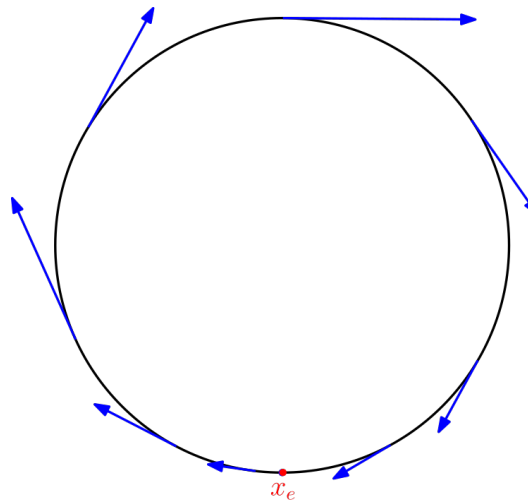
It’s extremely important to note that Lyapunov stability *does not* enforce convergence to the equilibrium point! In order for an equilibrium to be Lyapunov stable, the trajectories of the system simply need to *stay close* to it—convergence is a strictly stronger property. We say that an equilibrium  $x_e$  is *locally asymptotically stable* if it is Lyapunov stable and if starting in a region around  $x_e$  guarantees the trajectory will converge to  $x_e$ . It’s extremely important to note: *convergence does not imply Lyapunov stability*! It is a fundamental fact that not every system which converges is stable. In order to meet the conditions for asymptotic stability, one must have *both* Lyapunov stability and convergence. We present an intuitive counterexample in Figure 3.2.

Finally, *global* asymptotic stability is local asymptotic stability in which the region of convergence to the equilibrium is *all* of  $\mathbb{R}^n$ . An important fact regarding each of these stability definitions is that each form of stability is *preserved* under a change of coordinates.

<sup>1</sup> “That was hard.” - Blake Werner.



**Fig. 3.1** An equilibrium point  $x_e$  is *Lyapunov stable* if for any  $\epsilon > 0$  and  $t_0 \in \mathbb{R}$ , one can always find a  $\delta > 0$  such that starting in a ball of radius  $\delta$  around the equilibrium implies the system will *stay* in the ball of radius  $\epsilon$  around the equilibrium point. Notice: we're *not* required to stay in the ball of radius  $\delta$ ! All we ask is that we stay in the ball of radius  $\epsilon$ .



**Fig. 3.2** In this figure, the arrows represent a vector field tangent to the circle. For any initial condition, a curve tangent to the vector field will *converge* to  $x_e$ . However, if the curve starts to the left of  $x_e$ , it will *leave* any small neighborhood of  $x_e$ . Thus, the trajectories of the system defined by this vector field converge to  $x_e$  *but* do not satisfy the Lyapunov stability property.

**Proposition 3.3 (Stability is Preserved Under Change of Coordinates)** *Consider an unforced system with equilibrium point  $x_e \in \mathbb{R}^n$  and an invertible matrix  $T \in \mathbb{R}^{n \times n}$ . If  $x_e$  is (Lyapunov/asymptotically) stable, then  $Tx_e$  is (Lyapunov/asymptotically) stable for the transformed system defined by the change of state coordinates  $z = Tx$ .*

**Proof** See Problem 3.2. □

### 3.1.2 Stability of Linear, Time-Varying Systems

With these definitions in hand, we begin the task of characterizing the stability of equilibria of linear, time-varying systems. In particular, we'll analyze the stability of the equilibrium  $x_e = 0$  of the unforced systems,

$$\dot{x} = A(t)x(t), \quad x[k+1] = A[k]x[k]. \quad (3.10)$$

Recall that, for an initial condition  $x_0 \in \mathbb{R}^n$ , the solutions to each state equation are written in terms of the state transition matrix as  $x(t) = \Phi(t, t_0)x_0$  (in the continuous-time case) and  $x[k] = \Phi[k, k_0]x_0$  (in the discrete-time case). In order to determine the stability of an equilibrium point, we must bound  $\|x(t) - x_e\|$ —for the case of  $x_e = 0$ , this reduced to a bound on  $\|x(t)\|$ . How might we construct this bound? A simple analysis reveals,

$$\|x(t)\| = \|\Phi(t, t_0)x_0\| \leq \|\Phi(t, t_0)\| \|x_0\|, \quad (3.11)$$

where we assume that we are using the induced operator norm on  $\Phi$ . Thus, we observe that, if we can control the size of  $\|\Phi(t, t_0)\|$ , we can control the size of  $\|x(t)\|$ . The following theorem uses this observation to make a necessary and sufficient characterization of stability in terms of the state transition matrix.

**Theorem 3.1 (State Transition Matrix Characterization of Stability)** *Consider an unforced (continuous or discrete-time) linear system with state transition matrix  $\Phi : \mathbf{T} \rightarrow \mathbb{R}^{n \times n}$ ,  $\mathbf{T} = \{(t_1, t_0) \in \mathcal{T} : t_1 \geq t_0\}$ ,  $\mathcal{T} \in \{\mathbb{R}, \mathbb{Z}\}$ . The equilibrium  $x_e = 0$  is:*

1. Lyapunov stable if and only if for all  $t_0 \in \mathcal{T}$  there exists a  $K \geq 0$  for which,

$$\|\Phi(t, t_0)\| \leq K, \quad \forall t \in \mathcal{T}_{\geq t_0}. \quad (3.12)$$

2. Globally asymptotically stable if and only if for all  $t_0 \in \mathcal{T}$ ,

$$\lim_{t \rightarrow \infty} \|\Phi(t, t_0)\| = 0. \quad (3.13)$$

**Remark 3.4** Since condition (1) is necessary *and* sufficient for Lyapunov stability, violation of condition (1) automatically implies instability of  $x_e = 0$ . Thus, Theorem 3.1 also characterizes the *instability* of linear, time-varying systems.

**Remark 3.5** This theorem simultaneously deals with the continuous-time ( $\mathcal{T} = \mathbb{R}$ ) and discrete-time ( $\mathcal{T} = \mathbb{Z}$ ) cases. As such,  $\Phi(t, t_0)$  refers to either the continuous or discrete-time state transition matrix, and  $t, t_0$  refer to times in  $\mathcal{T}$ , not just  $\mathbb{R}$ .

**Proof** In this proof, we'll assume without loss of generality that the matrix norm on  $\Phi$  is the induced operator norm. We'll prove the discrete and continuous cases simultaneously, with  $t$  and  $t_0$  representing variables in  $\mathcal{T}$  (not necessarily  $\mathbb{R}$ ).

First, let's prove item (1). Suppose that, for all  $t_0 \in \mathcal{T}$ , there exists a  $K > 0$  for which  $\|\Phi(t, t_0)\| \leq K$  for all  $t \geq t_0$ . We wish to show that  $x_e = 0$  is *stable*. Fix an  $\epsilon > 0$  and a  $t_0 \in \mathcal{T}$ . Then, for  $t \geq t_0$  and  $x_0 \in \mathbb{R}^n$ ,

$$\|\varphi(t, t_0, x_0)\| = \|\Phi(t, t_0)x_0\| \leq \|\Phi(t, t_0)\| \|x_0\| \leq K \|x_0\|. \quad (3.14)$$

Picking  $\delta = \epsilon/K$ , it follows that  $\|x_0\| < \delta$  implies

$$\|\varphi(t, t_0, x_0)\| < K\delta = \epsilon, \quad \forall t \geq t_0. \quad (3.15)$$

We conclude that the equilibrium  $x_e = 0$  is Lyapunov stable. Now, we show the other direction. Suppose the equilibrium  $x_e = 0$  is Lyapunov stable. Then, for all  $\epsilon > 0$  and  $t_0 \in \mathcal{T}$ , there exists a  $\delta(t_0, \epsilon) > 0$  for which

$$\|x_0\| < \delta(t_0, \epsilon) \implies \|\varphi(t, t_0, x_0)\| < \epsilon. \quad (3.16)$$

Fix a  $t_0 \in \mathcal{T}$  and an  $\epsilon > 0$ . Assuming  $x_0 \neq 0$  and  $\|x_0\| < \delta(t_0, \epsilon)$ , one has,

$$\|\varphi(t, t_0, x_0)\| < \epsilon \quad (3.17)$$

$$\|\Phi(t, t_0)x_0\| < \epsilon \quad (3.18)$$

$$\|x_0\| \left\| \Phi(t, t_0) \frac{x_0}{\|x_0\|} \right\| < \epsilon. \quad (3.19)$$

Now, we take the supremum of both sides, subject to  $\|x_0\| < \delta(t_0, \epsilon)$  and  $x_0 \neq 0$ . Under this supremum, the strict inequality is replaced by a non-strict inequality, the  $\|x_0\|$  term is replaced by  $\delta(t_0, \epsilon)$ , and the  $\Phi$  term is replaced by the operator norm of  $\Phi(t, t_0)$  (since the operator norm is computed  $\|\Phi(t, t_0)\| = \sup_{\|u\|=1} \|\Phi(t, t_0)u\|$ ). This yields,

$$\sup_{\|x_0\| < \delta, x_0 \neq 0} \|x_0\| \left\| \Phi(t, t_0) \frac{x_0}{\|x_0\|} \right\| \leq \epsilon \quad (3.20)$$

$$\delta(\epsilon) \|\Phi(t, t_0)\| \leq \epsilon \quad (3.21)$$

$$\|\Phi(t, t_0)\| \leq \frac{\epsilon}{\delta(\epsilon)}. \quad (3.22)$$

Since this bound holds for all  $t \geq t_0$ , we conclude that the reverse direction of (1) holds.

Now, we prove item (2). We want to show  $x_e = 0$  is globally asymptotically stable if and only if  $\lim_{t \rightarrow \infty} \|\Phi(t, t_0)\| = 0$  for all  $t_0 \in \mathcal{T}$ . First, suppose that for all  $t_0 \in \mathcal{T}$ ,  $\lim_{t \rightarrow \infty} \|\Phi(t, t_0)\| = 0$ . In order to show global asymptotic stability, we must show that  $x_e = 0$  is stable in the sense of Lyapunov and that solutions limit to zero from all initial conditions. First, we'll show stability in the sense of Lyapunov. Fix a time  $t_0 \in \mathcal{T}$ . Since  $\lim_{t \rightarrow \infty} \|\Phi(t, t_0)\| = 0$ , one has that  $\sup_{t \geq t_0} \|\Phi(t, t_0)\| < \infty$  (this is a property of continuous functions in the continuous-time case and a property of convergent sequences for the discrete-time case). Defining  $K = \sup_{t \geq t_0} \|\Phi(t, t_0)\|$ , we conclude from item (1) that  $x_e = 0$  must be stable. Now, we show the limiting property. For any  $t_0 \in \mathcal{T}$  and  $x_0 \in \mathbb{R}^n$ , we have,

$$0 \leq \|\varphi(t, t_0, x_0)\| \leq \|\Phi(t, t_0)\| \|x_0\| \quad (3.23)$$

$$0 \leq \lim_{t \rightarrow \infty} \|\varphi(t, t_0, x_0)\| \leq \lim_{t \rightarrow \infty} \|\Phi(t, t_0)\| \|x_0\| = 0. \quad (3.24)$$

Since  $x_e$  is Lyapunov stable and satisfies the limiting property for all initial conditions, we conclude that  $x_e = 0$  is globally asymptotically stable. Now, we proceed in the other direction. Suppose  $x_e = 0$  is globally asymptotically stable. Then, for all  $t_0 \in \mathcal{T}$  and  $x_0 \in \mathbb{R}^n$ ,

$$\lim_{t \rightarrow \infty} \|\Phi(t, t_0)x_0\| = 0. \quad (3.25)$$

We wish to show that for all  $t_0 \in \mathcal{T}$ ,  $\lim_{t \rightarrow \infty} \|\Phi(t, t_0)\| = 0$ . We'll prove this property by picking some clever  $x_0$  vectors. Fix a time  $t_0 \in \mathcal{T}$ . For each standard basis vector  $e_i \in \mathbb{R}^n$  (containing a 1 in index  $i$  and zeros elsewhere),  $\lim_{t \rightarrow \infty} \|\Phi(t, t_0)e_i\| = 0$ . In other words, the  $i$ 'th column of  $\Phi(t, t_0)$  converges to the zero vector as  $t \rightarrow \infty$ . It then follows that  $\lim_{t \rightarrow \infty} \Phi(t, t_0) = 0$ , which in turn implies  $\lim_{t \rightarrow \infty} \|\Phi(t, t_0)\| = 0$ .  $\square$

As a corollary of this result, we show that, in the continuous-time case, one only needs to check the validity of items (1) and (2) for the special case of  $t_0 = 0$ .

**Corollary 3.1 (Continuous-Time Stability From  $t_0 = 0$ )** *Consider an unforced, continuous-time linear, time-varying system with state transition matrix  $\Phi : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}^{n \times n}$ . Then, the equilibrium  $x_e = 0$  is:*

1. Lyapunov stable if and only if there exists a  $K \geq 0$  for which,

$$\|\Phi(t, 0)\| \leq K, \quad \forall t \in \mathbb{R}_{\geq 0}. \quad (3.26)$$

2. Globally asymptotically stable if and only if,

$$\lim_{t \rightarrow \infty} \|\Phi(t, 0)\| = 0. \quad (3.27)$$

**Exercise 3.4** Using the composition property of the continuous-time state transition matrix, provide a proof of Corollary 3.1.

**Exercise 3.5** Show that a discrete-time analogue of Corollary 3.1 holds if  $\det A[k] \neq 0$  for all  $k \in \mathbb{Z}$ . What goes wrong in the case where  $A[k]$  is singular?

These results tell us that, in both the discrete and continuous-time cases, the Lyapunov and global asymptotic stability of the equilibrium  $x_e = 0$  is entirely characterized by the norm of the state transition matrix. What about local asymptotic stability? What about other equilibria? The following proposition offers the answers: local and global asymptotic stability are *equivalent* for linear systems, and  $x_e = 0$  is the only possible asymptotically stable equilibrium.

**Proposition 3.4 (Local Asymptotic Stability & Nonzero Equilibria)** *Consider an unforced (continuous or discrete-time) linear system.*

1. Global is equivalent to local: the equilibrium  $x_e = 0$  is globally asymptotically stable if and only if it is locally asymptotically stable.
2. Zero is the unique asymptotically stable equilibrium:  $x_e = 0$  is the only possible (locally or globally) asymptotically stable equilibrium of a linear system.



*Remark 3.6* Remember—these properties are only true for linear systems! Both items of the proposition are *false* for general, nonlinear systems.

**Proof** First, we'll prove item (1). One direction of this proof is easy - global asymptotic stability automatically implies local asymptotic stability. For the other direction, suppose  $x_e = 0$  is locally asymptotically stable. Fix a time  $t_0 \in \mathcal{T}$ . Then, there exists an  $\epsilon_0 > 0$  for which  $\|x_0\| < \epsilon_0$  implies  $\lim_{t \rightarrow \infty} \|\Phi(t, t_0)x_0\| = 0$ . This implies that, for  $v_i = e_i/(\epsilon_0 + \epsilon_1)$ , where  $\epsilon_1$  is an arbitrary, positive constant,  $\lim_{t \rightarrow \infty} \|\Phi(t, t_0)v_i\| = 0$ . Thus, we conclude that each column of  $\Phi(t, t_0)$  converges to zero as  $t \rightarrow \infty$ , which implies  $\lim_{t \rightarrow \infty} \|\Phi(t, t_0)\| = 0$ . Since this is true for all  $t_0 \in \mathcal{T}$ , we conclude by Theorem 3.1 that the equilibrium  $x_e = 0$  is globally asymptotically stable.

Now, prove item (2). Here, we'll prove the continuous-time case—the discrete-time case is left as an exercise below. Suppose  $x_e$  is a globally asymptotically stable equilibrium point. Performing the change of coordinates to  $z = x - x_e$ , we have,

$$\dot{z}(t) = \dot{x}(t) = A(t)x = A(t)(z(t) + x_e) = A(t)z(t) + A(t)x_e. \quad (3.28)$$

Let's solve this state equation from an arbitrary initial condition  $z_0 \in \mathbb{R}^n$ . For  $\Phi$  the state transition matrix of the original system, we have,

$$z(t) = \Phi(t, t_0)z_0 + \left( \int_{t_0}^t \Phi(t, \tau)A(\tau)x_e d\tau \right). \quad (3.29)$$

Since  $x_e$  is an equilibrium point of  $\dot{x}(t) = A(t)x(t)$ , it follows from the definition of piecewise continuity that  $A(\tau)x_e = 0$  for all but a finite number of points in the interval  $[t_0, t]$ . Thus, the integral term vanishes, which leaves us with

$$z(t) = \Phi(t, t_0)z_0. \quad (3.30)$$

Since  $x_e$  is a globally asymptotically stable equilibrium point of  $\dot{x}(t) = A(t)x(t)$ ,  $z_e = x_e - x_e = 0$  must be a globally asymptotically stable equilibrium point of  $\dot{z}(t) = A(t)z(t) + A(t)x_e$ . Thus, for all  $t_0 \in \mathbb{R}$  and  $z_0 \in \mathbb{R}^n$ ,  $z(t) = \Phi(t, t_0)z_0$  converges to 0 as  $t \rightarrow \infty$ . This implies that  $\lim_{t \rightarrow \infty} \Phi(t, t_0) = 0$ . But, Theorem 3.1 tells us this implies  $x = 0$  is a globally asymptotically stable equilibrium point of  $\dot{x}(t) = A(t)x(t)$ . Since both  $x_e$  and 0 are globally asymptotically stable equilibria, it follows that  $x_e = 0$ . The discrete-time case follows similarly.  $\square$

**Exercise 3.6** Complete the proof of Proposition 3.4 by showing item (2) also holds in the discrete-time case.

This tells us that, at best, all we can hope for is that a nonzero equilibrium point be stable in the sense of Lyapunov. As a consequence of this result, we'll focus on the equilibrium point  $x_e = 0$  going forward, as we'll find much more interesting behavior there.

### 3.1.3 Stability of Linear, Time-Invariant Systems

Let's recap what we've done so far. So far, we've stated precise definitions of Lyapunov and asymptotic stability, and have some up with a complete characterization of stability in the linear, time-varying case. In the past, once we've developed a comprehensive linear, time-varying theory, we've specialized to the linear, time-invariant case. In this case, we often find

the additional structure afforded by time-invariance yields stronger results. Now, we once again replicate this process for stability. What might we like out of a linear, time-invariant stability theory? Let's take a look at the linear, time-invariant state transition matrices:

$$\dot{x}(t) = Ax(t) \implies \Phi(t, t_0) = \exp(A(t - t_0)) \quad (3.31)$$

$$x[k + 1] = Ax[k] \implies \Phi[k, k_0] = A^{k-k_0}. \quad (3.32)$$

The structure of the state transition matrix as an *exponential* in the continuous-time case and a *matrix power* in the discrete-time case suggest that we can get much stronger stability properties under the assumption of time-invariance. In particular, instead of asking that our equilibria fall under the blanket of *asymptotic* stability, perhaps we can request a particular mode of convergence! The matrix exponential seems to suggest that *exponential* convergence to the equilibrium is possible in the continuous-time case, while the matrix power seems to suggest that *geometric* convergence is possible in the discrete-time case. By convention, we group both of these convergence modes under the common label of *exponential stability*. For the sake of generality, we define exponential stability for an arbitrary (not necessarily time-invariant) system.

**Definition 3.3 (Exponential Stability)** Consider an unforced system with state transition map  $\varphi : \mathbf{T} \times \mathbb{R}^n \rightarrow \mathbb{R}^n$  and equilibrium point  $x_e \in \mathbb{R}^n$ .

1. Local exponential stability (LES): if  $\mathcal{T} = \mathbb{R}$ ,  $x_e$  is locally exponentially stable if, for all  $t_0 \in \mathbb{R}$ , there exist constants  $\epsilon_0 > 0$ ,  $M \geq 0$ , and  $\alpha > 0$  for which,

$$\|x_0 - x_e\| < \epsilon_0 \implies \|\varphi(t, t_0, x_0) - x_e\| \leq Me^{-\alpha(t-t_0)} \|x_0 - x_e\|, \quad \forall t \geq t_0. \quad (3.33)$$

2. Local exponential stability (LES): if  $\mathcal{T} = \mathbb{Z}$ ,  $x_e$  is locally exponentially stable if, for all  $k_0 \in \mathbb{Z}$ , there exist constants  $\epsilon_0 > 0$ ,  $M \geq 0$ , and  $\alpha \in [0, 1)$  for which,

$$\|x_0 - x_e\| < \epsilon_0 \implies \|\varphi(k, k_0, x_0) - x_e\| \leq M\alpha^{k-k_0} \|x_0 - x_e\|, \quad \forall k \geq k_0. \quad (3.34)$$

3. Global exponential stability (GES): if  $\mathcal{T} = \mathbb{R}$ ,  $x_e$  is globally exponentially stable if, for all  $t_0 \in \mathbb{R}$ , there exist constants  $M \geq 0$  and  $\alpha > 0$  such that for all  $x_0 \in \mathbb{R}^n$ ,

$$\|\varphi(t, t_0, x_0) - x_e\| \leq Me^{-\alpha(t-t_0)} \|x_0 - x_e\|, \quad \forall t \geq t_0. \quad (3.35)$$

4. Global exponential stability (GES): if  $\mathcal{T} = \mathbb{Z}$ ,  $x_e$  is globally exponentially stable if, for all  $k_0 \in \mathbb{Z}$ , there exist constants  $M \geq 0$  and  $\alpha \in [0, 1)$  such that for all  $x_0 \in \mathbb{R}^n$ ,

$$\|\varphi(k, k_0, x_0) - x_e\| \leq M\alpha^{k-k_0} \|x_0 - x_e\|, \quad \forall k \geq k_0. \quad (3.36)$$

*Remark 3.7* It's important to note that exponential stability enforces a *stronger condition* than SISL plus exponential convergence. Instead of asking that the equilibrium be SISL and that trajectories converge exponentially to the equilibrium point, we provide a single bound,  $Me^{-\alpha(t-t_0)} \|x_0 - x_e\|$  (or  $\alpha^{k-k_0}$  in discrete-time), which accomplishes both goals. This structure enforces that the upper bound on  $\|\varphi(t, t_0, x_0) - x_e\|$  be linear in  $\|x_0 - x_e\|$ , a constraint which would not be present had we assumed an arbitrary exponential mode of convergence.

*Remark 3.8* All too often in engineering, computer science, and mathematics, we focus on the mode of convergence, and neglect the presence of any constants associated to the mode

of convergence. This tendency is reflected in the definition of exponential stability, where we request that we get exponential convergence of the form  $Me^{-\alpha(t-t_0)}$  (or  $M\alpha^{k-k_0}$ ) for *some*  $M \geq 0$ . Although the exponential convergence certainly dominates as  $t \rightarrow \infty$ , the constant  $M$  should not be ignored in practice! If one is working with a system with tight state constraints, exponential convergence as  $t \rightarrow \infty$  might not be relevant if  $M$  is overly large—if the state constraints are violated for small  $t$ , the asymptotic behavior is irrelevant! In practice, it pays to come up with bounds on the size of  $M$  to confirm that the constants associated with the convergence are not of a dramatic order of magnitude. This helps ensure that both the short and long-term behaviors of the system are of desirable character.

How does exponential stability relate to asymptotic and Lyapunov stability? An initial guess is that exponential is a strictly *stronger* condition than both asymptotic and Lyapunov. The following proposition confirms that this is in fact the case. As its proof is straightforward, we leave the details to the reader.

**Proposition 3.5 (Hierarchy of Stability Modes)** *Consider an unforced system with state transition map  $\varphi : \mathbf{T} \times \mathbb{R}^n \rightarrow \mathbb{R}^n$  and equilibrium point  $x_e \in \mathbb{R}^n$ . The modes of stability satisfy the following hierarchy:*

$$\begin{array}{c} \text{(Local) Exponential Stability} \\ \Downarrow \\ \text{(Local) Asymptotic Stability} \\ \Downarrow \\ \text{Lyapunov Stability.} \end{array}$$

*Remark 3.9* Close inspection of the definition of Lyapunov stability will confirm that Lyapunov stability is *always* local - hence, no additional “local” moniker is needed.

**Exercise 3.7** Provide a proof of Proposition 3.5.

Recall that, in the LTI case, *time elapsed* is the relevant quantity, rather than objective start and end times. As a consequence of this property, it follows that, for an LTI system, it is sufficient to verify that stability properties hold for an initial time  $t_0 = 0$  in order to make claims about stability properties holding for all time.

**Proposition 3.6 (LTI Stability at  $t_0 = 0$ )** *Consider an unforced LTI system with state transition map  $\varphi : \mathbf{T} \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ . An equilibrium point  $x_e \in \mathbb{R}^n$  is (Lyapunov/asymptotically/exponentially) stable if and only the conditions for (Lyapunov/asymptotical/exponential) stability are satisfied for the initial time  $t_0 = 0$ .*

**Exercise 3.8** Prove Proposition 3.6.

What this proposition states is that, if we wish to verify the stability of a time-invariant system, we only need to check that the requisite conditions are met for  $t_0 = 0$ . For the case of a time-invariant continuous-time system, for instance, to prove global exponential stability it is sufficient to exhibit constants  $M \geq 0$  and  $\alpha > 0$  for which

$$\|\varphi(t, 0, x_0) - x_e\| \leq Me^{-\alpha t} \|x_0 - x_e\|, \quad \forall t \geq 0, x_0 \in \mathbb{R}^n. \quad (3.37)$$

That is, we *no longer* need to produce constants for each initial time  $t_0 \in \mathbb{R}$ —showing there exist constants for  $t_0 = 0$  is sufficient to conclude exponential stability. With this simplifying fact in mind, we begin the process of characterizing the stability of LTI systems.

### 3.1.3.1 The Continuous-Time Case

Let's take stock of where we are. Earlier, we demonstrated that every locally asymptotically LTV stable is in fact globally asymptotically stable, and that the only equilibrium with asymptotic behavior is  $x_e = 0$ . Additionally, we introduced the concept of *exponential stability*, inspired by the presence of the matrix exponential. With these concepts in mind, a reasonable goal in the LTI case is to characterize the global exponential stability of the equilibrium  $x_e = 0$ .

Let's do a little bit of basic analysis to see how we might bound  $\|x(t)\|$  (the distance from the solution  $x(\cdot)$  at time  $t$  to the zero equilibrium) in the continuous-time, LTI case. Let  $x(t)$  denote the solution to the initial value problem,  $\dot{x} = Ax$ ,  $x(0) = x_0$ , at time  $t$ . One has,

$$\|x(t)\| = \|\exp(At)x_0\| \leq \|\exp(At)\| \|x_0\|, \quad \forall t \geq 0. \quad (3.38)$$

Thus, we encounter a structure familiar from the linear, time-varying case—one in which we bound the norm of the system's trajectories by separately bounding the norms of the transition matrix and the initial condition. Since, for exponential stability, we'd like an upper bound of the form  $Me^{-\alpha t} \|x_0\|$ ,  $\forall t \geq 0$ , the structure above suggests that we identify  $M$  and  $\alpha$  for which

$$\|\exp(At)\| \leq Me^{-\alpha t} \quad \forall t \geq 0. \quad (3.39)$$

How should we pick  $M$  and  $\alpha$ ? For an arbitrary, unstructured matrix  $A \in \mathbb{R}^{n \times n}$ , the answer to this question is obscured by the complex form of the matrix exponential. In order to prove there exist  $M$  and  $\alpha$  satisfying the conditions for exponential stability, we therefore transform the  $A$  matrix into a form with a simpler exponential. Recall that we can compute the exponential of any matrix in Jordan canonical form,  $\text{blkdiag}(J_1, \dots, J_p)$ ,  $J_i = \lambda_i I + N_0$ , in closed form, and that any matrix can be transformed into a Jordan canonical form. Thus, as a starting point, we'll work on identifying an upper bound on the norm of the exponential of a single Jordan block,  $J = \lambda I + N_0$ . From there, we'll subsequently identify bounds on  $\|\exp(\text{blkdiag}(J_1, \dots, J_p)t)\|$  and  $\|\exp(At)\|$ . Using these bounds, we'll come up with sharp characterizations of global exponential stability. Let's get started!

Recall that the exponential of a Jordan block  $J = \lambda I + N_0$  is computed,

$$\exp(Jt) = e^{\lambda t} \begin{bmatrix} 1 & t & \dots & \frac{t^{n-1}}{(n-1)!} \\ 0 & 1 & \ddots & \vdots \\ \vdots & & \ddots & t \\ 0 & \dots & \dots & 1 \end{bmatrix}. \quad (3.40)$$

We note that the matrix is preceded by an exponential term,  $e^{\lambda t}$ , and is filled with polynomial terms. Thus, as a first step, we'll show that we can bound the norm of  $\exp(Jt)$  by a product of an exponential and a polynomial.

**Lemma 3.1 (Exponential-Polynomial Bound on Norm of Jordan block)** *Consider an  $n \times n$  Jordan block  $J = \lambda I + N_0 \in \mathbb{C}^{n \times n}$ . For any norm  $\|\cdot\|$  on  $\mathbb{C}^{n \times n}$ , there exist constants  $c_0, \dots, c_{n-1} \in \mathbb{R}_{\geq 0}$  for which,*

$$\|\exp(Jt)\| \leq e^{\text{Re}(\lambda)t} (c_0 + c_1 t + \dots c_{n-1} t^{n-1}), \quad \forall t \geq 0. \quad (3.41)$$

**Proof** This proof will proceed as follows. First, we'll come up with a norm on  $\mathbb{C}^{n \times n}$  that makes the desired inequality easy to prove. Then, we'll use norm equivalence on  $\mathbb{C}^{n \times n}$  to show that the inequality can be produced under any matrix norm.

First, we come up with a simple, easy-to-compute norm on  $\mathbb{C}^{n \times n}$ . For a matrix  $A \in \mathbb{C}^{n \times n}$ , consider the norm  $\|A\| = \sum_{i,j=1}^n |a_{ij}|$ , which sums the magnitudes of all of the entries of  $A$  (this is the same as stretching  $A$  out into a vector and taking its  $\ell_1$  norm). Examining the structure of the Jordan block, it's clear that for all  $t \geq 0$ ,

$$\|\exp(Jt)\| \leq |e^{\lambda t}|(\hat{c}_0 + \hat{c}_1 t + \dots + \hat{c}_{n-1} t^{n-1}) \quad (3.42)$$

$$= e^{\operatorname{Re}(\lambda)t}(\hat{c}_0 + \hat{c}_1 t + \dots + \hat{c}_{n-1} t^{n-1}), \quad (3.43)$$

for some nonnegative, real constants  $\hat{c}_0, \dots, \hat{c}_{n-1}$ . Now, take any norm  $\|\cdot\|'$  on  $\mathbb{C}^{n \times n}$ . Since  $\mathbb{C}^{n \times n}$  is a finite dimensional vector space, we conclude from norm equivalence between  $\|\cdot\|$  and  $\|\cdot\|'$  that there exist constants  $c_0, \dots, c_{n-1} \in \mathbb{R}$ , which yield the desired bound.  $\square$

**Exercise 3.9** Confirm that  $\|A\| = \sum_{i,j=1}^n |a_{ij}|$  defines a norm on  $\mathbb{C}^{n \times n}$ .

Great! This produces a bound on the norm of the exponential in terms of a polynomial multiplied by the exponential of  $\operatorname{Re}(\lambda)t$ . In order to characterize exponential stability, however, we must eliminate the polynomial term. We accomplish this by showing we can bound a polynomial by a constant times an exponential of an arbitrary positive rate.

**Lemma 3.2 (Polynomials are Bounded by Exponentials)** *Consider a polynomial with real coefficients,  $p_n(t) = a_n t^n + \dots + a_1 t + a_0$ ,  $a_i \in \mathbb{R}$ ,  $t \in \mathbb{R}$ . For any  $\epsilon > 0$ , there exists a constant  $K > 0$  such that, for all  $t \geq 0$ ,  $p_n(t) \leq K e^{\epsilon t}$ .*

**Remark 3.10** The key aspect of this lemma is that we have *complete control* over the choice of  $\epsilon > 0$ . Since we can pick  $\epsilon$  to be as small as we want, we can bound  $e^{\operatorname{Re}(\lambda)t}(c_0 + \dots + c_{n-1} t^{n-1})$  by a single exponential without dramatically changing the rate of the exponential.

**Proof** Without loss of generality, assume that the coefficients of the polynomial are non-negative (if this is not the case, take their magnitude to form an upper bound on  $p_n(t)$ ). The proof of this result follows from the Taylor series definition of  $\exp(\epsilon t)$ . Fix an  $\epsilon > 0$ . The Taylor series of  $\exp(\epsilon t)$  is computed,

$$\exp(\epsilon t) = \sum_{k=0}^{\infty} \frac{(\epsilon t)^k}{k!} = 1 + \epsilon t + \frac{\epsilon^2 t^2}{2!} + \dots + \frac{\epsilon^n t^n}{n!} + \sum_{k=n+1}^{\infty} \frac{\epsilon^k t^k}{k!}. \quad (3.44)$$

Thus, for all  $t \geq 0$ ,

$$1 + \epsilon t + \frac{\epsilon^2 t^2}{2!} + \dots + \frac{\epsilon^n t^n}{n!} \leq \exp(\epsilon t). \quad (3.45)$$

In order to bound  $p_n(t) = a_n t^n + \dots + a_1 t + a_0$  by the product of a constant and  $\exp(\epsilon t)$ , it's therefore sufficient to bound  $p_n(t)$  by a constant times the  $n$ 'th order polynomial above. Choose a constant  $K \geq 0$  such that  $K(\epsilon^k/k!) \geq a_k$ , for all  $k \in 0, \dots, n$ . Since  $0 \leq a_k \leq K(\epsilon^k/k!)$ , it then follows that

$$p_n(t) \leq K(1 + \epsilon t + \frac{\epsilon^2 t^2}{2!} + \dots + \frac{\epsilon^n t^n}{n!}) \leq K \exp(\epsilon t), \quad \forall t \geq 0, \quad (3.46)$$

This completes the proof.  $\square$

We stress—the critical component of this result is that we can choose *any* positive exponential rate  $\epsilon$ —we are not constrained by a given value of  $\epsilon$  for a given polynomial. This fact will be essential in determining non-conservative conditions for exponential stability. With this result in hand, we now have all of the pieces we need to fully characterize the global exponential stability of continuous-time, LTI systems.

**Theorem 3.2 (Eigenvalue Characterization of Exponential Stability)** *Consider the unforced continuous-time, LTI system  $\dot{x}(t) = Ax(t)$ . The equilibrium  $x_e = 0$  of the system is globally exponentially stable if and only if, for all eigenvalues  $\lambda$  of  $A$ ,  $\text{Re}(\lambda) < 0$ .*

**Proof** First, we'll show sufficiency of the eigenvalue condition. Suppose all eigenvalues  $\lambda$  of  $A$  satisfy  $\text{Re}(\lambda) < 0$ . Proposition 3.6 tells us that, in order to confirm global exponential stability of  $x_e = 0$ , it's sufficient to show that there exist constants  $M \geq 0$ ,  $\alpha > 0$  for which

$$\|\varphi(t, 0, x_0)\| \leq Me^{-\alpha t} \|x_0\|, \quad \forall t \geq 0, x_0 \in \mathbb{R}^n. \quad (3.47)$$

That is, we only need to exhibit valid constants for the initial time  $t_0 = 0$ . First, let's transform into Jordan form. We know there exists an invertible matrix  $T \in \mathbb{C}^{n \times n}$  for which  $A = TJT^{-1}$ , where  $J = \text{blkdiag}(J_1, \dots, J_p)$  is in Jordan canonical form with Jordan blocks  $J_1, \dots, J_p$ . Fix an initial condition  $x_0 \in \mathbb{R}^n$ . The norm of  $\varphi(t, 0, x_0)$  is bounded in terms of the exponential of  $J$  as,

$$\|\varphi(t, 0, x_0)\| = \|\exp(At)x_0\| = \|T \exp(Jt)T^{-1}x_0\| \leq \|T\| \|\exp(Jt)\| \|T^{-1}\| \|x_0\|, \quad (3.48)$$

for all  $t \geq 0$ . Now, we use Lemma 3.1 to bound the norm of each Jordan block. For each  $i \in 1, \dots, p$ , there exist constants  $c_{j,i} \geq 0, j \in 1, \dots, n-1$ , for which

$$\|\exp(J_i t)\| \leq e^{\text{Re}(\lambda_i)t} (c_{0,i} + c_{1,i}t + \dots + c_{n-1,i}t^{n-1}), \quad \forall t \geq 0, \quad (3.49)$$

where we've picked the degree of the polynomial to be  $n-1$  without loss of generality (since  $A \in \mathbb{R}^{n \times n}$ , the size of  $J_i$  is limited to  $n \times n$ , which bounds the degree of the polynomial by  $n-1$ ). Define  $c_j = \max_i c_{j,i}$  and  $\lambda = \max_i \text{Re}(\lambda_i)$ . For such  $c_j$  and  $\lambda$ , there exists a constant  $C \geq 0$  (accounting for the matrix norm and the number of Jordan blocks) for which

$$\|\exp(Jt)\| \leq Ce^{\lambda t} (c_0 + c_1 t + \dots + c_{n-1} t^{n-1}), \quad \forall t \geq 0. \quad (3.50)$$

Since  $\text{Re}(\lambda_i) < 0$  for all  $i$ , it follows that  $\lambda < 0$ . Now, pick  $\epsilon > 0$  such that  $\lambda + \epsilon < 0$ . For such an  $\epsilon$ , Lemma 3.2 tells us there exists a  $K \geq 0$  for which  $c_0 + \dots + c_{n-1} t^{n-1} \leq Ke^{\epsilon t}$ , for all  $t \geq 0$ . Substituting this upper bound into the inequality above, we find,

$$\|\exp(Jt)\| \leq Ce^{\lambda t} Ke^{\epsilon t} = CK e^{(\lambda+\epsilon)t}, \quad \forall t \geq 0. \quad (3.51)$$

Substituting back into our original bound on  $\|\varphi(t, 0, x_0)\|$ , it follows that

$$\|\varphi(t, 0, x_0)\| \leq \|T\| \|T^{-1}\| CK e^{(\lambda+\epsilon)t} \|x_0\|, \quad \forall t \geq 0. \quad (3.52)$$

As  $x_0$  was arbitrary, this bound holds for all  $x_0 \in \mathbb{R}^n$ . Since  $(\lambda + \epsilon) < 0$ , we conclude that the equilibrium  $x_e = 0$  is globally exponentially stable.

Now, we'll show necessity of the eigenvalue condition. Suppose the system is globally exponentially stable. We will show that, for all eigenvalues  $\lambda$  of  $A$ ,  $\text{Re}(\lambda) < 0$ . Suppose for contradiction that there exists an eigenvalue  $\lambda$  of  $A$  for which  $\text{Re}(\lambda) \geq 0$ . If  $J = T^{-1}AT$  is

a Jordan canonical form of  $A$ , it follows that  $\lim_{t \rightarrow \infty} \|\exp(Jt)\| \neq 0$ . Thus,

$$\lim_{t \rightarrow \infty} \|T \exp(Jt) T^{-1}\| = \lim_{t \rightarrow \infty} \|\exp(At)\| \neq 0. \quad (3.53)$$

Since  $\exp(At) = \Phi(t, 0)$ , we conclude that  $\lim_{t \rightarrow \infty} \|\Phi(t, 0)\| \neq 0$ . This, by Theorem 3.1 means that  $x_e = 0$  is not asymptotically stable and is therefore not globally exponentially stable. Contradiction! We conclude that  $\operatorname{Re}(\lambda) < 0$  for all eigenvalues  $\lambda$  of  $A$ .  $\square$

*Remark 3.11* A quick remark is in order about the proof of necessity of the eigenvalue condition. In order to reach a contradiction, why didn't we choose an eigenvector  $v$  corresponding to the eigenvalue  $\lambda$  and show that  $\lim_{t \rightarrow \infty} \exp(At)v = e^{\lambda t}v \neq 0$ ? The trouble with this argument is that, in general, an eigenvalue  $\lambda$  of  $A$  can be complex. For such eigenvalues, the corresponding eigenvectors will also be complex! In order to reach a contradiction using this method, we must exhibit a *real* vector  $v \in \mathbb{R}^n$  for which  $\lim_{t \rightarrow \infty} \exp(At)v \neq 0$ , not a complex vector! A valid, alternate proof strategy would be to separate the eigenvector  $v \in \mathbb{C}^n$  into its real and complex components, and perform an analysis of each individual component. In the proof above, we avoided this subtlety by working directly with the norm of the exponential.

This result tells us that the global exponential stability of the equilibrium  $x_e = 0$  of an LTI system,  $\dot{x}(t) = Ax(t)$  is *entirely* determined by the eigenvalues of  $A$ . This result has enormous implications for linear control design, as we'll see across the remainder of the course. Since matrices with  $\operatorname{Re}(\lambda) < 0$  play such a fundamental role, we give them a special name.

**Definition 3.4 (Open Left Half-Plane/Hurwitz Matrix)** The set  $\{\lambda \in \mathbb{C} : \operatorname{Re}(\lambda) < 0\}$  is called the *open left half plane*. A matrix  $A \in \mathbb{C}^{n \times n}$  is said to be *Hurwitz* if all of its eigenvalues belong to the open left half plane.

Note that Hurwitz matrices are also referred to as *stable* matrices. Using this language, we say that the equilibrium  $x_e = 0$  of  $\dot{x}(t) = Ax(t)$  is globally exponentially stable if and only if all of the eigenvalues of  $A$  belong to the open left half plane, or if and only if  $A$  is Hurwitz.

Recall that in the general linear, time-varying case, we showed that  $x_e = 0$  is locally asymptotically stable if and only if it is globally asymptotically stable, and that  $x_e = 0$  is the *only* asymptotically stable equilibrium point. Now, we confirm that analogous results hold for the linear, time-invariant case. Additionally, we will show that asymptotic and exponential stability are equivalent for linear, time-invariant systems.

**Proposition 3.7 (Stability Modes of Continuous-Time LTI Systems)** Consider an unforced continuous-time, LTI system  $\dot{x}(t) = Ax(t)$ . The following properties hold:

1. Local is equivalent to global: the equilibrium  $x_e = 0$  is locally exponentially stable if and only if it is globally exponentially stable.
2. Asymptotic is equivalent to exponential:  $x_e = 0$  is (locally) asymptotically stable if and only if it is (locally) exponentially stable.
3. Zero is the unique exponentially stable equilibrium:  $x_e = 0$  is the only possible (locally or globally) exponentially stable equilibrium point of the system.

**Proof** We start with item (1). One direction is obvious—global exponential stability automatically implies local exponential stability. Now, we prove the other direction. Suppose  $x_e = 0$  is locally exponentially stable. Then, there exist constants  $\epsilon_0, M, \alpha > 0$  for which

$$\|\varphi(t, 0, x_0)\| \leq Me^{-\alpha t} \|x_0\|, \quad \forall t \geq 0, \|x_0\| < \epsilon_0. \quad (3.54)$$

Now, fix an arbitrary, nonzero initial condition  $x_0 \in \mathbb{R}^n$ . Then, there exists an  $\epsilon_1 > 0$  for which  $x'_0 = x_0/(\epsilon_0 + \epsilon_1)$  satisfies  $\|x'_0\| < \epsilon_0$ . For such an  $x'_0$ , one has,

$$\|\varphi(t, 0, x'_0)\| \leq Me^{-\alpha t} \|x'_0\| \quad (3.55)$$

$$(\epsilon_0 + \epsilon_1) \|\varphi(t, 0, x'_0)\| \leq Me^{-\alpha t} \|x_0\| \quad (3.56)$$

$$\|\varphi(t, 0, x_0)\| \leq Me^{-\alpha t} \|x_0\|, \quad (3.57)$$

where in the last step, we use that the state transition map is linear in the initial condition. Since this holds for all  $x_0 \in \mathbb{R}^n$ , we conclude that  $x_e = 0$  is globally exponentially stable.

Next, we prove item (2). Since we just proved that global and local exponential stability are equivalent, it's sufficient to prove item (2) for the case of global exponential stability. Again, one direction is simple—we already know that global exponential stability implies global asymptotic stability. Now, we show the other direction. Suppose  $x_e = 0$  is globally asymptotically stable. Then, one has,

$$\lim_{t \rightarrow \infty} \|\Phi(t, 0)\| = \lim_{t \rightarrow \infty} \|\exp(At)\| = 0. \quad (3.58)$$

Suppose  $A = TJT^{-1}$ , where  $J$  is a Jordan canonical form of  $A$ . Then, the above implies  $\lim_{t \rightarrow \infty} \|T \exp(Jt) T^{-1}\| = 0$ , which in turn implies  $\lim_{t \rightarrow \infty} \|\exp(Jt)\| = 0$ . In order for  $\exp(Jt)$  to converge to zero, the exponential of each Jordan block must converge to zero. Thus, for each eigenvalue  $\lambda$  of  $A$  and corresponding Jordan block  $J_\lambda = \lambda I + N_0 \in \mathbb{C}^{m \times m}$ , we must have,

$$\lim_{t \rightarrow \infty} |e^{\lambda t}| \left\| \begin{bmatrix} 1 & t & \dots & \frac{t^{m-1}}{(m-1)!} \\ 0 & 1 & \ddots & \vdots \\ \vdots & & \ddots & t \\ 0 & \dots & \dots & 1 \end{bmatrix} \right\| = \lim_{t \rightarrow \infty} e^{\operatorname{Re}(\lambda)t} \left\| \begin{bmatrix} 1 & t & \dots & \frac{t^{m-1}}{(m-1)!} \\ 0 & 1 & \ddots & \vdots \\ \vdots & & \ddots & t \\ 0 & \dots & \dots & 1 \end{bmatrix} \right\| = 0. \quad (3.59)$$

A necessary condition for this limit to hold is that  $\operatorname{Re}(\lambda) < 0$ . Since  $\operatorname{Re}(\lambda) < 0$  for all eigenvalues  $\lambda$  of  $A$  implies global exponential stability, we conclude that  $x_e = 0$  must be globally exponentially stable.

Finally, we prove item (3). Suppose there exists an equilibrium point  $x_e \in \mathbb{R}^n$  which is (locally or globally) exponentially stable. Then, such an equilibrium is (locally or globally) asymptotically stable, which by Proposition 3.4 implies  $x_e = 0$ .  $\square$

The eigenvalue value characterization of stability for linear, time-invariant systems yields a tempting but *false* characterization of stability of linear, time-varying systems—one might guess that  $\dot{x}(t) = A(t)x(t)$  is stable if  $\operatorname{Re}(\lambda(A(t))) < 0$  for all  $t$ . However, this is *not* the case! In fact, a linear, time-varying system satisfying this condition might even be *unstable*. The following example illustrates exactly this.

*Example 3.1 (Eigenvalues do not determine stability of LTV systems)* Consider the linear dynamical system  $\dot{x}(t) = A(t)x(t)$ , where  $A(t)$  is defined,

$$A(t) = \begin{bmatrix} -1 + a \cos^2(t) & 1 - a \cos(t) \sin(t) \\ -1 + a \cos(t) \sin(t) & -1 + a \sin^2(t) \end{bmatrix}, \quad (3.60)$$



for  $a \in (1, 2)$  is some fixed constant. One may check that the eigenvalues of  $A(t)$  satisfy  $\text{Re}(\lambda_i) = -(2 - \lambda)/2$ , and that the state transition matrix is computed,

$$\Phi(t, 0) = \begin{bmatrix} e^{(a-1)t} \cos(t) & e^{-t} \sin(t) \\ -e^{(a-1)t} \sin(t) & e^{-t} \cos(t) \end{bmatrix}. \quad (3.61)$$

If  $a \in (1, 2)$ , the norm of the state transition matrix is *not* bounded! So, the system must be unstable, despite having eigenvalues satisfying the LTI exponential stability condition. As a consequence of this example, we note that *not only* do eigenvalues not confirm exponential stability of a general LTV system, but they do not even confirm *stability*.

**Exercise 3.10** Verify the calculation of the eigenvalues and state transition matrix of  $A(t)$  in the example above.

Thus far, we've characterized the exponential and asymptotic stability of LTI systems. What about standard Lyapunov stability? What about instability? Consider the following result, which provides conditions for both the Lyapunov stability and instability of continuous-time, LTI systems.

**Theorem 3.3 (Lyapunov Stability of LTI Systems)** *The equilibrium  $x_e = 0$  of an unforced continuous-time, LTI system  $\dot{x} = Ax$  is Lyapunov stable if and only if the following two conditions are met:*

1. *Eigenvalues:* For all eigenvalues  $\lambda$  of  $A$ ,  $\text{Re}(\lambda) \leq 0$ .
2. *Multiplicity:* The algebraic and geometric multiplicities of all eigenvalues  $\lambda$  of  $A$  satisfying  $\text{Re}(\lambda) = 0$  are equal.

*Remark 3.12* Since this result is a necessary *and* sufficient statement, if either condition of Theorem 3.3 is violated we can automatically conclude that  $x_e = 0$  is unstable.

**Proof** See Problem 3.1 for the details, and Example 3.2 for a hint. □

*Example 3.2 (Unstable LTI System)* Using Theorem 3.3, let's produce an example of an unstable LTI system with all zero eigenvalues. Consider the continuous-time, LTI system,

$$\dot{x}(t) = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} x(t). \quad (3.62)$$

We recognize the  $A$  matrix as a  $2 \times 2$  Jordan block for which  $\lambda = 0$ . Since there is a 1 above the diagonal, it *must be* that the geometric and algebraic multiplicities of  $\lambda = 0$  do not match (if they did match, there would be a basis of eigenvectors for the eigenspace of  $\lambda = 0$  and the matrix would be diagonalizable). Computing the exponential of  $At$ , we find,

$$\exp(At) = \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix}. \quad (3.63)$$

This clearly has an unbounded norm. From Theorem 3.1, we conclude that the equilibrium  $x_e = 0$  is unstable. This is also made evident by taking the initial condition  $x_0 = [0, 1]^\top$ , for which the solution tends to infinity as  $t \rightarrow \infty$ .

### 3.1.3.2 The Discrete-Time Case

Now, we repeat our study of linear, time-invariant exponential stability for the discrete-time case. In the continuous-time case, we started by examining the norm of the state transition matrix for an initial time  $t_0 = 0$ . In discrete-time, we'll do the same! Recall that for the discrete-time, LTI system  $x[k+1] = Ax[k]$ , the state transition matrix is computed  $\Phi[k, k_0] = A^{k-k_0}$ ,  $k \geq k_0$ . We therefore have,

$$\|\varphi(k, 0, x_0)\| = \|\Phi[k, 0]x_0\| \leq \|A^k\| \|x_0\|. \quad (3.64)$$

Thus, in order to certify the exponential stability of a discrete-time system, it's sufficient to show that  $\|A^k\| \leq M\alpha^k$ , for some  $M \geq 0$  and  $\alpha \in [0, 1)$ . To find a sharp set of conditions for which this occurs, we'll once again begin with the case in which  $A$  is a single Jordan block. Recall that the exponent of a Jordan block  $J = \lambda I + N_0 \in \mathbb{C}^{n \times n}$  is computed,

$$J^k = \sum_{j=0}^k \binom{k}{j} \lambda^{k-j} N_0^j = \begin{bmatrix} \lambda^k \binom{k}{1} \lambda^{k-1} \binom{k}{2} \lambda^{k-2} \dots & & & \\ 0 & \lambda^k & \binom{k}{1} \lambda^{k-1} \dots & \\ \vdots & \vdots & \ddots & \ddots \\ 0 & 0 & \dots & \lambda^k \binom{k}{1} \lambda^{k-1} \\ 0 & 0 & \dots & 0 & \lambda^k \end{bmatrix}. \quad (3.65)$$

How might we come up with a bound on the norm of  $J^k$ ? Based on the structure of  $J^k$ , and our development in the continuous-time case, it seems like we should seek a bound on the norm of  $J^k$  using  $|\lambda|^k$ . In order to develop an effective bound, we first need the following, intermediate result.

**Lemma 3.3 (Bounding Combinations with Exponents)** *Let  $\lambda \geq 0$ . For any  $\epsilon > 1$  and any integer  $r \geq 0$ , there exists an  $M \geq 0$  for which*

$$\binom{k}{r} \lambda^k \leq M(\epsilon\lambda)^k, \quad \forall k \in \mathbb{Z}_{\geq 0}. \quad (3.66)$$

**Proof** Fix an integer  $r \geq 0$ . In the case of  $\lambda = 0$ , the result is obvious, since both sides of the proposed inequality are zero. Thus, we assume  $\lambda > 0$ . Recall that  $k$  choose  $r$  is computed  $k!/(r!(k-r)!)$ . Let's work on bounding this quantity. We have,

$$\frac{k!}{(k-r)!} = \prod_{j=0}^{r-1} (k-j) \leq k^r \implies \binom{k}{r} = \frac{k!}{r!(k-r)!} \leq \frac{k^r}{r!}. \quad (3.67)$$

We conclude that for fixed  $r$ ,  $k$  choose  $r$  has polynomial growth. It follows from the Taylor series definition of the exponent  $e^k$  that, for any integer  $r \geq 0$  and any  $\epsilon > 1$ , there exists a  $C \geq 0$  for which  $k^r \leq C\epsilon^k$ , for all  $k \geq 0$  (i.e. polynomial growth is bounded by exponential growth). Let's apply this fact to the problem at hand. Fixing an  $\epsilon > 1$ , we have,

$$\binom{k}{r} \lambda^k \leq \frac{k^r}{r!} \lambda^k \leq \frac{C}{r!} \epsilon^k \lambda^k = \frac{C}{r!} (\epsilon\lambda)^k. \quad (3.68)$$

Defining  $M = (C/r!)$ , the result follows.  $\square$

Using this result, we can come up with the desired bound on  $\|J^k\|$  for a Jordan block  $J$ .

**Lemma 3.4 (Bounding a Jordan Block Exponent)** *Consider a Jordan block for a nonzero eigenvalue,  $J = \lambda I + N_0 \in \mathbb{C}^{n \times n}$ ,  $\lambda \neq 0$ . For any norm  $\|\cdot\|$  on  $\mathbb{C}^{n \times n}$  and any  $\epsilon > 1$ , there exists an  $M \geq 0$  for which*

$$\|J^k\| \leq M|\epsilon\lambda|^k, \quad \forall k \geq 0. \quad (3.69)$$

**Proof** Fix a norm  $\|\cdot\|$  on  $\mathbb{C}^{n \times n}$ . For any  $k \geq 0$ ,

$$\|J^k\| = \left\| \sum_{j=0}^k \binom{k}{j} \lambda^{k-j} N_0^j \right\| \leq \sum_{j=0}^k \binom{k}{j} |\lambda|^{k-j} \|N_0^j\|. \quad (3.70)$$

Now, we'd like to make an upper bound that holds for all  $k \geq 0$ . We know that, for  $j \geq n$ ,  $N_0^j = 0$ . This tells us that, for all integers  $k \geq 0$ ,

$$\|J^k\| \leq \sum_{j=0}^k \binom{k}{j} |\lambda|^{k-j} \|N_0^j\| \quad (3.71)$$

$$= \sum_{j=0}^{n-1} \binom{k}{j} |\lambda|^{k-j} \|N_0^j\| \quad (3.72)$$

$$= \binom{k}{0} |\lambda|^k + \binom{k}{1} |\lambda|^{k-1} \|N_0\| + \dots + \binom{k}{n-1} |\lambda|^{k-n+1} \|N_0^{n-1}\| \quad (3.73)$$

$$= \binom{k}{0} |\lambda|^k + \|N_0\| |\lambda|^{-1} \binom{k}{1} |\lambda|^k + \dots + \|N_0^{n-1}\| |\lambda|^{-(n-1)} \binom{k}{n-1} |\lambda|^k. \quad (3.74)$$

Now, fix  $\epsilon > 1$ . By the previous lemma, there exist  $M_0, \dots, M_{n-1} > 0$  for which,

$$\|J^k\| \leq M_0 |\epsilon\lambda|^k + M_1 |\lambda|^{-1} |\epsilon\lambda|^k + \dots + M_{n-1} |\lambda|^{-(n-1)} |\epsilon\lambda|^k \leq M |\epsilon\lambda|^k, \quad (3.75)$$

where the final constant  $M$  is taken from adding  $M_0, \dots, M_{n-1}$ . For  $k < n$ , this bound also holds. Thus, for all  $\lambda \neq 0$ , the result follows.  $\square$

Putting these results together, we arrive at a necessary and sufficient characterization of stability for the discrete-time, LTI system.

**Theorem 3.4 (Discrete-Time Eigenvalue Characterization of Stability)** *Consider an unforced discrete-time, LTI system  $x[k+1] = Ax[k]$ . The equilibrium  $x_e = 0$  of the system is globally exponentially stable if and only if, for all eigenvalues  $\lambda$  of  $A$ ,  $|\lambda| < 1$ .*

**Proof** First, we'll show that  $|\lambda| < 1$  is sufficient for global exponential stability. In the case where every eigenvalue of  $A$  equals zero, it must be that  $A^m = 0$  for some  $m > 0$ , from which global exponential stability follows. Now, consider the case where  $A$  has at least one nonzero eigenvalue. For any  $x_0 \in \mathbb{R}^n$ , one has

$$\|x[k]\| = \|A^k x_0\| \leq \|A^k\| \|x_0\| \leq \|T\| \|J^k\| \|T^{-1}\| \|x_0\|, \quad (3.76)$$

where  $T$  is a transformation taking  $A$  into the Jordan canonical form  $J = \text{blkdiag}(J_1, \dots, J_p)$ . Now, let  $\alpha = \max_{i=1, \dots, p} |\lambda_i|$ . Since  $|\lambda| < 1$  for all eigenvalues  $\lambda$  of  $A$  and not all eigenvalues of  $A$  are zero, it follows that  $\alpha \in (0, 1)$ . The previous lemma implies that, for all  $\epsilon > 1$ , there exists an  $M > 0$  such that,

$$\|J^k\| \leq M(\epsilon\alpha)^k, \quad \forall k \in \mathbb{Z}, \quad (3.77)$$

since the powers of all Jordan blocks corresponding to zero eigenvalues disappear after  $k \geq n$ . Selecting an  $\epsilon$  for which  $\alpha\epsilon \in (0, 1)$ , we have,

$$\|x[k]\| \leq \|T\| \|T^{-1}\| M(\epsilon\alpha)^k \|x_0\|, \quad \forall k \geq 0, \quad (3.78)$$

which implies the global exponential stability of  $x_e = 0$ .

Now, we show that  $|\lambda| < 1$  is necessary for global exponential stability. Suppose  $x_e = 0$  is globally exponentially stable. For contradiction, suppose there exists an eigenvalue  $\lambda$  of  $A$  for which  $\|\lambda\| \geq 1$ . In such an event, it's clear (via the Jordan decomposition) that  $\|\Phi[k, 0]\| = \|A^k\|$  does not limit to zero as  $k \rightarrow \infty$ . Thus, by Theorem 3.1,  $x_e = 0$  cannot be globally asymptotically stable, and therefore cannot be globally exponentially stable.  $\square$

Just as we gave the set of exponentially stable, continuous-time  $A$  matrices a special name, we give the set of exponentially stable, discrete-time  $A$  matrices a special name.

**Definition 3.5 (Open Unit Disk/Schur Matrix)** The set  $\{\lambda \in \mathbb{C} : |\lambda| < 1\}$  is called the open unit disk in the complex plane. A matrix  $A$  is said to be *Schur* if all of its eigenvalues belong to the open unit disk.

Thus, a *Hurwitz* matrix is one that yields an exponentially stable continuous-time system, and a *Schur* matrix is one that yields an exponentially stable discrete-time system. Now, we continue to replicate our steps from the continuous-time case. Next, we'll find that the zero equilibrium of an exponentially stable, discrete-time system possesses the same stability properties as the zero equilibrium of an exponentially stable, continuous-time system. Since the proofs are similar to the continuous-time case, we leave the details to the reader.

**Proposition 3.8 (Stability Modes of Discrete-Time LTI Systems)** Consider a discrete-time, LTI system  $x[k+1] = Ax[k]$ . The following properties hold:

1. Local is equivalent to global: the equilibrium  $x_e = 0$  is locally exponentially stable if and only if it is globally exponentially stable.
2. Asymptotic is equivalent to exponential:  $x_e = 0$  is (locally) asymptotically stable if and only if it is (locally) exponentially stable.
3. Zero is the unique exponentially stable equilibrium:  $x_e = 0$  is the only possible (locally or globally) exponentially stable equilibrium of the system.

**Exercise 3.11** Prove Proposition 3.8.

Finally, we characterize the Lyapunov stability of the equilibria of discrete-time systems.

**Theorem 3.5 (Lyapunov Stability of LTI Systems)** The equilibrium  $x_e = 0$  of an unforced discrete-time, LTI system  $x[k+1] = Ax[k]$  is Lyapunov stable if and only if the following two conditions are met:

1. *Eigenvalues:* For all eigenvalues  $\lambda$  of  $A$ ,  $|\lambda| \leq 1$ .
2. *Multiplicity:* The algebraic and geometric multiplicities of all eigenvalues  $\lambda$  of  $A$  satisfying  $|\lambda| = 1$  are equal.

**Proof** See Problem 3.1.  $\square$

### 3.1.4 Further Reading

The treatment of stability in this chapter is mainly inspired by [8] and [25] - both texts take an approach based on the state transition matrix similar to those which we present above. Example 3.1, of an unstable linear, time-varying system, is from [25].

### 3.1.5 Problems

**Problem 3.1 (Lyapunov Stability of LTI Systems [25])** In this problem, we'll characterize the Lyapunov stability of LTI systems.

1. Prove that an eigenvalue  $\lambda$  of  $A$  has a diagonal Jordan block if its algebraic multiplicity equals its geometric multiplicity.
2. Using your answer to part (1), write a proof of Proposition 3.3.
3. Using your answer to part (1), write a proof of Proposition 3.5.

**Problem 3.2 (Stability & System Relations)** In this problem, we'll examine how the stability of systems is preserved under coordinate transforms. Here, we'll consider an arbitrary (continuous or discrete-time) input-free system with state transition map  $\varphi : \mathbf{T} \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ , where  $\mathbf{T} = \{(t_1, t_0) \in \mathcal{T} \times \mathcal{T} : t_1 \geq t_0\}$ .

1. Let  $\hat{\varphi} : \mathbf{T} \times \mathbb{R}^n \rightarrow \mathbb{R}^n$  be the state transition map of a second, input-free system on  $\mathbb{R}^n$ . Let  $T : \mathbb{R}^n \rightarrow \mathbb{R}^n$  be a linear map. The systems  $\varphi$  and  $\hat{\varphi}$  are said to be  $T$ -related if,

$$T(\varphi(t, t_0, x_0)) = \hat{\varphi}(t, t_0, T(x_0)), \quad \forall t \geq t_0 \in \mathcal{T}, x_0 \in \mathbb{R}^m. \quad (3.79)$$

If the systems have the form  $\dot{x} = A(t)x$  and  $\dot{\hat{x}} = \hat{A}(t)\hat{x}$ , find sufficient conditions on  $A$  and  $\hat{A}$  such that the two systems are  $T$ -related.

2. Prove that if  $T$  is invertible, then the equilibrium  $x_e = 0$  of  $\dot{x}(t) = A(t)x(t)$  is (Lyapunov/asymptotically/exponentially) stable if and only if the equilibrium  $\hat{x}_e = 0$  of  $\dot{\hat{x}}(t) = \hat{A}(t)\hat{x}(t)$  is (Lyapunov/asymptotically/exponentially) stable.
3. Suppose now that  $T$  is a surjective linear mapping from  $\mathbb{R}^n \rightarrow \mathbb{R}^k$ . What can you conclude about the stability of  $\hat{x}_e = 0$  from the stability of  $x_e = 0$ ? What about the case where  $T$  is injective? Back up your claims with proofs or counterexamples.

**Problem 3.3 (A Simple Stability Condition [9])** Consider the unforced linear, time-varying system  $\dot{x}(t) = A(t)x(t)$ , where  $A(\cdot) \in PC(\mathbb{R}, \mathbb{R}^{n \times n})$ . Show that if  $A(t)$  is *skew-symmetric* for all  $t \in \mathbb{R}$  ( $A(t) = -A^\top(t)$ ), then  $x_e = 0$  is Lyapunov stable.

**Problem 3.4 (Robustness of Exponential Stability ★)** In this problem, we'll show that exponential stability is *robust* under small perturbations.

1. (**Hard**—you can skip this subproblem if you can't find a solution after giving it some thought) Consider a family of polynomials parameterized by  $t$ ,

$$f(s, t) = a_n(t)s^n + \dots + a_1(t)s + a_0(t), \quad (3.80)$$

where each  $a_i : \mathbb{R} \rightarrow \mathbb{R}$  is continuous. Prove there exist continuous functions  $\lambda_i : \mathbb{R} \rightarrow \mathbb{C}$ ,  $i = 1, \dots, n$ , such that for all  $t_0 \in \mathbb{R}$ , each  $\lambda_i(t_0)$  corresponds to a root of  $f(s, t_0)$ .

2. Prove there exists a continuous function  $\text{spec} : \mathbb{R}^{n \times n} \rightarrow \mathbb{C}^n$ , mapping a matrix  $A \in \mathbb{R}^{n \times n}$  to a vector containing its eigenvalues.
3. Let  $A \in \mathbb{R}^{n \times n}$ . Consider the perturbed systems,

$$\dot{x}(t) = (A + \Delta)x(t), \quad x[k+1] = (A + \Delta)x[k], \quad (3.81)$$

where  $\Delta \in \mathbb{R}^{n \times n}$ . Suppose each system is globally exponentially stable for  $\Delta = 0$ . In each case, prove there exists an  $M > 0$  such that for all  $\Delta : \|\Delta\| < M$ , the system remains globally exponentially stable.

**Problem 3.5 (Separating Hyperplane for a Linear Dynamical System [6])** Let  $c \in \mathbb{R}^n$  be a nonzero vector. The *hyperplane* passing through 0 defined by  $c$  is the set,

$$\mathcal{H}_c = \{x \in \mathbb{R}^n : c^\top x = 0\} \subseteq \mathbb{R}^n. \quad (3.82)$$

Consider a continuous-time, LTI system  $\dot{x}(t) = Ax(t)$ , where  $A \in \mathbb{R}^{n \times n}$  and  $x \in \mathbb{R}^n$ . A hyperplane  $\mathcal{H}_c$  passing through zero is said to be a *separating hyperplane* for this system if no trajectory of the system ever crosses the hyperplane. That is, if  $c^\top \varphi(t, t_0, x_0) < 0$  for some  $t \in \mathbb{R}$ , it is impossible to have  $c^\top \varphi(t', t_0, x_0) > 0$  for another time  $t' \in \mathbb{R}$ . Assuming the eigenvalues of  $A$  are all distinct, explain how to find *all* separating hyperplanes of  $\dot{x}(t) = Ax(t)$ . Find the conditions on  $A$  under which there are *no* separating hyperplanes.

## 3.2 Invariance-Based Approaches to Stability

Previously, we studied the stability of linear, time-varying and linear, time-invariant systems through an *analytical* lens. In particular, we developed characterizations of Lyapunov, asymptotic, and exponential stability by establishing bounds on the deviation of trajectories from an equilibrium point. In this section, we'll take a more *qualitative* approach to studying stability. We'll primarily focus on the case of linear, time-invariant systems, and see how to interpret the behavior of a system as a function of its eigenvalues. Our key tool in this endeavor will be *invariance*. We'll prove the invariant subspace theorem, which tells us about special, invariant subspaces of the state space, and will look at Lyapunov theorems, which harness the invariance of certain distinguished subsets of the state space to make conclusions about stability. Let's begin!

### 3.2.1 The Real Jordan Form

We'll begin our qualitative study of stability by examining the effect of the eigenvalues of the  $A$  matrix on the behavior of the system. In order to completely characterize this relationship, we'll first need to develop a little bit more Jordan form theory. Recall that in the previous section, we analyzed systems of the form  $\dot{x}(t) = Ax(t)$  and  $x[k+1] = Ax[k]$  by transforming  $A \in \mathbb{R}^{n \times n}$  into a *potentially complex* Jordan form,  $J \in \mathbb{C}^{n \times n}$ . Once we had this Jordan form, we were able to compute the matrix exponential in closed form and develop the bounds we needed to study stability.

Although this transformation was convenient for the sake of analysis, its complex nature obscures much of the behavior of the original, real system.<sup>2</sup> For instance, if we take the following  $A$  matrix into its Jordan form,

$$A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \implies J = \begin{bmatrix} j & 0 \\ 0 & -j \end{bmatrix}, \quad (3.83)$$

we immediately lose an intuitive understanding of what each component of the solution represents. In order to gain an intuitive understanding of systems governed by real matrices, we'd therefore like to develop a purely real analogue of the Jordan form. Like the "standard" Jordan form, we should be able to easily compute the exponential of a matrix in "real" Jordan form, and should be able to transform any real, square matrix into a "real" Jordan form. What, then, should this form be? In order to motivate the structure of the real Jordan form, we recall the following fact about the eigenvalues of real matrices.

**Fact** Complex eigenvalues of real matrices come in *conjugate pairs*. If  $\lambda = \sigma + j\omega \in \mathbb{C}$  is an eigenvalue of  $A \in \mathbb{R}^{n \times n}$ , then its conjugate,  $\bar{\lambda} = \sigma - j\omega \in \mathbb{C}$ , is also an eigenvalue of  $A$ .  $\square$

**Exercise 3.12** Provide a proof of the fact above. *Hint: the characteristic polynomial of a real matrix has real coefficients.*

Instead of using Jordan blocks of the form  $J = \lambda I + N_0$ , which result in complex matrices for  $\lambda \in \mathbb{C}$ , perhaps we could construct Jordan-like blocks which encode the real and complex

<sup>2</sup> To quote [17], "In general it is not polite to hand someone a complex solution to a real system of differential equations."

parts of  $\lambda$ . Since complex eigenvalues come in *pairs* for real matrices, maybe we could deal with both  $\lambda \in \mathbb{C}$  and  $\bar{\lambda} \in \mathbb{C}$  with a single block! The following structure—termed the *complex Jordan block*—does exactly this.

**Definition 3.6 (Complex Jordan Block)** The complex Jordan block of size  $2n \times 2n$  corresponding to  $\lambda = \sigma + j\omega \in \mathbb{C}$ ,  $\sigma, \omega \in \mathbb{R}$ , is the matrix

$$J_{\sigma,\omega} = \begin{bmatrix} \Lambda I_2 & 0 & \dots & 0 \\ 0 & \Lambda I_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \Lambda I_2 \\ 0 & 0 & \dots & 0 & \Lambda \end{bmatrix} \in \mathbb{R}^{2n \times 2n}, \text{ where } \Lambda = \begin{bmatrix} \sigma & -\omega \\ \omega & \sigma \end{bmatrix} \text{ and } I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \quad (3.84)$$

*Remark 3.13* Despite its name, the complex Jordan block is an *entirely real* matrix! The name *complex Jordan block* follows from the fact that it corresponds to a complex eigenvalue.

The structure of the complex Jordan block closely mirrors that of the standard Jordan block. Here, we trade the eigenvalues  $\lambda \in \mathbb{C}$  down the diagonal for  $2 \times 2$  blocks  $\Lambda \in \mathbb{R}^{2 \times 2}$  down the diagonal, where  $\Lambda$  encodes  $\lambda$  and its conjugate. Likewise, we trade the ones on the superdiagonal for a set of  $2 \times 2$  identity matrices just above the block diagonal. For instance, in the event where one has two blocks of  $\Lambda$ , the complex Jordan block would be,

$$J_{\sigma,\omega} = \begin{bmatrix} \Lambda & I_2 \\ 0 & \Lambda \end{bmatrix} = \begin{bmatrix} \sigma & -\omega & 1 & 0 \\ \omega & \sigma & 0 & 1 \\ 0 & 0 & \sigma & -\omega \\ 0 & 0 & \omega & \sigma \end{bmatrix} \in \mathbb{R}^{4 \times 4}. \quad (3.85)$$

Above, when laying out the problem of the real Jordan form, we stated that there are two conditions we'd like to meet. First, we'd like for the exponential of the real Jordan form to be readily computable. Secondly, we'd like to be able to transform *any* real matrix into a real Jordan form. Now, we complete the first item, and show that the complex Jordan block behaves well under exponentiation. Since the proof of the following proposition is similar to the proof of the exponential of a standard Jordan block, we leave it as an exercise.

**Proposition 3.9 (Exponential of a Complex Jordan Block)** Let  $J_{\sigma,\omega}$  be a complex Jordan block of size  $2n \times 2n$  corresponding to eigenvalue  $\lambda = \sigma + j\omega$ . The exponential of  $J_{\sigma,\omega}t$  is computed,

$$\exp(J_{\sigma,\omega}t) = e^{\sigma t} \begin{bmatrix} R(t) & tR(t) & \dots & \frac{t^{n-1}}{(n-1)!}R(t) \\ 0 & R(t) & \ddots & \vdots \\ \vdots & & \ddots & tR(t) \\ 0 & \dots & \dots & R(t) \end{bmatrix}, \text{ where } R(t) = \begin{bmatrix} \cos(\omega t) & -\sin(\omega t) \\ \sin(\omega t) & \cos(\omega t) \end{bmatrix}. \quad (3.86)$$

**Exercise 3.13** Prove Proposition 3.9. See [31], Chapter 1.8, for a solution.

Thus, we observe that a complex Jordan block has a simple, easy to interpret exponential. We find that the real component,  $\sigma$ , entirely controls how fast the entries of the exponential *decay*, and that the complex component  $\omega$ , entirely controls how fast the entries of the exponential *oscillate*. These observations highlight the advantages of working with real numbers.



In  $\mathbb{R}^{n \times n}$ , it's much easier to interpret how the real and complex components of a given eigenvalue affect a system's behavior, since all functions in the exponential are real.

With the complex Jordan block in hand, we now have everything we need to define a purely real analogue of the Jordan form. The definition proceeds as follows: to every *strictly real* eigenvalue  $\lambda$  of  $A$ , one assigns a standard Jordan block  $J_\lambda = \lambda I + N_0$ , and to every *complex* eigenvalue  $\lambda = \sigma + j\omega$ , one assigns a complex Jordan block  $J_{\sigma, \omega}$ . Assembling each of these blocks into a block diagonal matrix yields a real analogue of the Jordan form.

**Definition 3.7 (Real Jordan Form)** Consider a matrix  $A \in \mathbb{R}^{n \times n}$  with  $p$  distinct, real eigenvalues and  $k$  distinct, conjugate pairs of complex eigenvalues,

$$\underbrace{\{\lambda_1, \dots, \lambda_p\}}_{\text{Real}}, \underbrace{\{\sigma_1 \pm j\omega_1, \dots, \sigma_k \pm j\omega_k\}}_{\text{Complex}} \subseteq \mathbb{C}. \quad (3.87)$$

A matrix  $J \in \mathbb{R}^{n \times n}$  is said to be a real Jordan form of  $A$  if it is of the form

$$J = \text{blkdiag}(J_1, \dots, J_r) \in \mathbb{R}^{n \times n}, \quad (3.88)$$

where each  $J_i$  is either a Jordan block of the form  $J_i = \lambda_i I + N_0$  (for a real eigenvalue  $\lambda_i$  of  $A$ ) or a complex Jordan block of the form  $J_i = J_{\sigma_i, \omega_i}$  (for a complex eigenvalue  $\sigma_i + j\omega_i$ ).

*Remark 3.14* Note that each complex Jordan block covers both  $\sigma + j\omega$  and  $\sigma - j\omega$ —there are *not* separate complex Jordan blocks for  $\sigma + j\omega$  and  $\sigma - j\omega$ . Each complex Jordan block is able to cover two eigenvalues since it itself is constructed from  $2 \times 2$  blocks.

Thus, a real Jordan form of a matrix  $A \in \mathbb{R}^{n \times n}$  is a block diagonal matrix where each real eigenvalue  $\lambda$  is assigned a standard Jordan block and each conjugate pair of complex eigenvalues,  $\sigma \pm j\omega$ , is assigned a complex Jordan block corresponding to  $\sigma + j\omega$ . The existence of conjugate pairs of complex eigenvalues necessitates that a complex Jordan block be composed of  $2 \times 2$  block matrices.

Now, we'd like to confirm that we can transform *any* real matrix  $A \in \mathbb{R}^{n \times n}$  into a real Jordan form. Fortunately, we've already done all the heavy lifting required to prove this result in the previous chapter. We'll find that, due to the close relationship between the real Jordan form and the standard Jordan form, transformations of a real matrix into a real Jordan form are constructed from the same basis of generalized eigenvectors as the standard Jordan form. As such, we don't need to reprove that a basis of generalized eigenvectors for our space exists—this was the trickiest step in constructing the standard Jordan form.

Let's get to work on proving this transformation exists. From the previous chapter, we already know how to compute the pieces of the transformation into real Jordan form associated to the real Jordan blocks,  $J = \lambda I + N_0$ . Now, we must develop a method to construct the pieces of the transformation associated to a complex Jordan block.

**Lemma 3.5 (Transformation into a Complex Jordan Block)** Consider a matrix  $A \in \mathbb{R}^{2n \times 2n}$  with two eigenvalues,  $\{\sigma \pm j\omega\}$ ,  $\sigma, \omega \in \mathbb{R}$ . Suppose  $\{u_i + jv_i \in \mathbb{C}^{2n}, i = 1, \dots, n\}$  is the maximal chain of generalized eigenvectors of  $A$  for  $\sigma + j\omega$ . With respect to the basis  $\mathcal{B} = \{u_1, -v_1, \dots, u_n, -v_n\} \subseteq \mathbb{R}^{2n}$ ,  $[A]_{\mathcal{B}} = [u_1, -v_1, \dots, u_n, -v_n]^{-1} A [u_1, -v_1, \dots, u_n, -v_n]$  is a complex Jordan block for  $\sigma + j\omega$ .

**Proof** First, we show that the proposed basis  $\mathcal{B}$  is a basis for  $\mathbb{R}^{2n}$ . First, we show that if  $\{u_i + jv_i, i = 1, \dots, n\}$  is a maximal chain corresponding to  $\sigma + j\omega$ , then  $\{u_i - jv_i, i = 1, \dots, n\}$  is a maximal chain corresponding to  $\sigma - j\omega$ . We'll begin by establishing  $\{u_i - jv_i, i = 1, \dots, n\}$

is a chain for  $\sigma - j\omega$  using induction on  $i$ . Recall that the first element of a chain is always an eigenvector. For  $i = 1$ , we therefore have

$$A(u_1 + jv_1) = \lambda(u_1 + jv_1) \quad (3.89)$$

$$\overline{A(u_1 + jv_1)} = \overline{\lambda(u_1 + jv_1)} \quad (3.90)$$

$$A(u_1 - jv_1) = \bar{\lambda}(u_1 - jv_1), \quad (3.91)$$

where we drop the conjugate on  $A$  since it is real. Since the conjugate of  $\lambda$  is  $\bar{\lambda} = \sigma - j\omega$ , the result holds for  $i = 1$ . Now, we assume that the result holds for  $i - 1$ ,  $n > i > 1$ . Using the definition of a chain, we have,

$$(A - \lambda I)(u_i + jv_i) = u_{i-1} + jv_{i-1} \quad (3.92)$$

$$\overline{(A - \lambda I)(u_i + jv_i)} = \overline{u_{i-1} + jv_{i-1}} \quad (3.93)$$

$$(A - \bar{\lambda}I)(u_i - jv_i) = u_{i-1} - jv_{i-1}. \quad (3.94)$$

We conclude that  $u_i - jv_i$  is a member of the chain corresponding to  $\sigma - j\omega$ . Since chains corresponding to different generalized eigenvectors are independent, the set  $\{u_i + jv_i, u_i - jv_i, i = 1, \dots, n\}$  must form a basis for  $\mathbb{C}^{2n}$ . Since we cannot add another vector to the  $\{u_i - jv_i\}$  chain without the total number of vectors in this set exceeding  $2n$ , we conclude that the chain is maximal. Since the collection  $\{u_i + jv_i, u_i - jv_i, i = 1, \dots, n\} \subseteq \mathbb{C}^{2n}$  forms a basis for  $\mathbb{C}^{2n}$ , it must be that the collection  $\{u_i, -v_i, i = 1, \dots, n\} \subseteq \mathbb{R}^{2n}$  is linearly independent. Since this is a linearly independent collection of  $2n$  real vectors in  $\mathbb{R}^{2n}$ ,  $\mathcal{B} = \{u_i, -v_i, i = 1, \dots, n\}$  must form a basis for  $\mathbb{R}^{2n}$ . Now that we've shown that  $\mathcal{B}$  is indeed a basis, we return to the problem at hand. We want to show,

$$\begin{bmatrix} | & | & \dots & | & | \\ u_1 & -v_1 & \dots & u_n & -v_n \\ | & | & \dots & | & | \end{bmatrix} \begin{bmatrix} A I_2 & 0 & \dots & 0 \\ 0 & A & I_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & A & I_2 \\ 0 & 0 & \dots & 0 & A \end{bmatrix} = A \begin{bmatrix} | & | & \dots & | & | \\ u_1 & -v_1 & \dots & u_n & -v_n \\ | & | & \dots & | & | \end{bmatrix}, \quad A = \begin{bmatrix} \sigma & -\omega \\ \omega & \sigma \end{bmatrix}. \quad (3.95)$$

Let's go piece by piece. Looking at the first column of the left-hand side, we want to show,

$$[u_1 - v_1] A = A [u_1 - v_1] \quad (3.96)$$

$$[u_1 - v_1] \begin{bmatrix} \sigma & -\omega \\ \omega & \sigma \end{bmatrix} = A [u_1 - v_1] \quad (3.97)$$

$$[\sigma u_1 - \omega v_1 \quad -\omega u_1 - \sigma v_1] = A [u_1 - v_1], \quad (3.98)$$

Since  $u_1 + jv_1$  is the first element of a chain with eigenvalue  $\sigma + j\omega$ , it must be an eigenvector of  $A$  with eigenvalue  $\lambda = \sigma + j\omega$ . We therefore have  $A(u_1 + jv_1) = Au_1 + jAv_1 = (\sigma + j\omega)(u_1 + jv_1)$ . Expanding this product and equating real and complex parts, we have,

$$Au_1 + jAv_1 = \sigma u_1 + j\omega u_1 - \omega v_1 + j\sigma v_1 \quad (3.99)$$

$$\Rightarrow Au_1 = \sigma u_1 - \omega v_1 \quad (3.100)$$

$$\Rightarrow Av_1 = \omega u_1 + \sigma v_1. \quad (3.101)$$

This proves that the desired equality holds for  $i = 1$ . Now, let's show that we can get an equality for  $i > 1$ . First, we note that for  $i > 1$ ,

$$(A - \lambda I)(u_i + jv_i) = u_{i-1} + jv_{i-1}, \quad \lambda = \sigma + j\omega, \quad (3.102)$$

by definition of a chain of generalized eigenvectors. Expanding the left-hand side, we find

$$(A - (\sigma + j\omega)I)(u_i + jv_i) = A(u_i + jv_i) - \sigma u_i - \sigma jv_i - j\omega u_i + \omega v_i. \quad (3.103)$$

Equating the above with the right-hand side of the chain equality, it follows that

$$Au_i + jAv_i - \sigma u_i - \sigma jv_i - j\omega u_i + \omega v_i = u_{i-1} + jv_{i-1} \quad (3.104)$$

$$Au_i + jAv_i = \sigma u_i - \omega v_i + u_{i-1} + j(\omega u_i + \sigma v_i + v_{i-1}). \quad (3.105)$$

Equating real and complex parts, it follows that  $Au_i = \sigma u_i - \omega v_i + u_{i-1}$  and  $Av_i = \omega u_i + \sigma v_i + v_{i-1}$ , which implies

$$\begin{bmatrix} u_{i-1} & -v_{i-1} & u_i & -v_i \end{bmatrix} \begin{bmatrix} I_2 \\ A \end{bmatrix} = A \begin{bmatrix} u_i & -v_i \end{bmatrix}. \quad (3.106)$$

Thus, for  $i > 1$ , the required pattern holds. The result follows.  $\square$

With the use of Lemma 3.5, we can prove that we can transform *any* real matrix into a real Jordan form! In the following theorem, we show how to construct a transformation of any real matrix into a real Jordan form by use of its generalized eigenvectors.

**Theorem 3.6 (Transformation into Real Jordan Form)** *Consider a matrix  $A \in \mathbb{R}^{n \times n}$  with  $p$  distinct real eigenvalues and  $k$  conjugate pairs of distinct complex eigenvalues,*

$$\underbrace{\{\lambda_1, \dots, \lambda_p\}}_{\text{Real}}, \underbrace{\{\sigma_1 \pm j\omega_1, \dots, \sigma_k \pm j\omega_k\}}_{\text{Complex}} \subseteq \mathbb{C}. \quad (3.107)$$

*There exists a basis  $\mathcal{B}$  of  $\mathbb{R}^n$  consisting of vectors,*

$$\mathcal{B} = \underbrace{\{u_1, \dots, u_m\}}_{\text{Real } \lambda_i}, \underbrace{\{u_{m+1}, -v_{m+1}, \dots, u_{m+q}, -v_{m+q}\}}_{\text{Complex } \lambda_i} \subseteq \mathbb{R}^n, \quad (3.108)$$

*where each  $u_i$ ,  $i = 1, \dots, m$  is a generalized eigenvector corresponding to a real eigenvalue and each  $u_{m+i} + jv_{m+i}$ ,  $i = 1, \dots, q$  is a generalized eigenvector corresponding to a complex eigenvalue, for which  $[A]_{\mathcal{B}}$  is in real Jordan form.*

**Proof (Sketch)** We give a sketch of a proof of Theorem 3.6—a formal proof requires a little bit more bookkeeping. We know there exists a basis  $\mathcal{B}'$  for  $\mathbb{C}^n$  composed of chains of generalized eigenvectors of  $A$ —this is a key component of the “classic” Jordan form theorem which we discussed in the previous chapter. From this basis, add each generalized eigenvector corresponding to a strictly real eigenvalue to a new set,  $\mathcal{B}$ . Now, we consider the complex eigenvalues, which come in conjugate pairs  $\sigma_i \pm j\omega_i$ . Select the eigenvalue  $\sigma_i + j\omega_i$ , and all of its chains of generalized eigenvectors  $u_i + jv_i$ . Add each  $u_i$  and  $-v_i$  to  $\mathcal{B}$ , taking care to preserve the order of the chains. Application of Lemma 3.5 confirms that  $[A]_{\mathcal{B}}$  will indeed be in real Jordan form.  $\square$

**Exercise 3.14** Fill in the details required to make the proof of Theorem 3.6 fully rigorous.

**Exercise 3.15** A matrix  $J$  is given by,

$$J = \begin{bmatrix} -1 & 0 & 0 & 0 & 0 \\ 0 & -2 & 1 & 0 & 0 \\ 0 & -1 & -2 & 0 & 0 \\ 0 & 0 & 0 & -3 & 1 \\ 0 & 0 & 0 & 0 & -3 \end{bmatrix}. \quad (3.109)$$

Find the eigenvalues of  $J$  and  $\exp(Jt)$  by inspection.

We once again stress the principle:

*The real Jordan form is good for proofs and bad for numerical computation!*

Using the same examples as in Chapter 1, which showed that arbitrarily small perturbations can change the structure of a Jordan form, we find that the real Jordan form is also extremely numerically unstable! For instance, if we take the complex Jordan block,

$$A = \begin{bmatrix} \sigma - \omega & \\ \omega & \sigma \end{bmatrix}, \quad (3.110)$$

and *perturb* one of the entries by an arbitrarily small number  $\epsilon > 0$  to get,

$$A_\epsilon = \begin{bmatrix} \sigma - \omega + \epsilon & \\ \omega & \sigma \end{bmatrix}, \quad (3.111)$$

the resulting matrix is *no longer* a valid complex Jordan block! Thus, we emphasize that the real Jordan form—while an exceptional tool for conceptual mathematics—should be avoided for numerical computation.

### 3.2.2 Phase Portraits

Armed with the real Jordan form, we're now ready to begin our *qualitative* study of stability. We'll begin this study in the *planar* case, where  $\Sigma = \mathbb{R}^2$ . In this setting, we can easily sketch out the trajectories of systems and gain an intuition for the relationship between eigenvalues and the shape of trajectories. Further, the intuitions we develop in the planar case help in building an understanding of the behavior of general linear systems on  $\mathbb{R}^n$ .

As mentioned above, one reason that planar linear systems are appealing is that we can actually *sketch* the trajectories of linear systems on a pair of coordinate axes. For a planar system  $\dot{x} = Ax$  or  $x[k+1] = Ax[k]$ ,  $x \in \mathbb{R}^2$ , a *phase portrait* of the system is a plot in which the  $x$ -axis represents the  $x_1$  coordinate of the system and the  $y$ -axis represents the  $x_2$  coordinate of the system. In order to qualitatively understand the behavior of our system, we can plot  $(x_1(t), x_2(t))$  trajectories from a variety of initial conditions.

Our goal in this subsection is to completely *classify* all phase portraits of linear systems  $\dot{x} = Ax$  and  $x[k+1] = Ax[k]$ . This means that—up to an invertible change of coordinates on  $\mathbb{R}^2$ —we want to have an idea of what every possible phase portrait of a 2-D linear system looks like. This is where the real Jordan form comes in—as a result of Theorem 3.6, we know that *any* matrix  $A \in \mathbb{R}^{2 \times 2}$  can be transformed into one of the following,

$$\begin{bmatrix} \lambda & 0 \\ 0 & \mu \end{bmatrix}, \begin{bmatrix} \lambda & 1 \\ 0 & \lambda \end{bmatrix}, \begin{bmatrix} \sigma & -\omega \\ \omega & \sigma \end{bmatrix}, \quad (3.112)$$

where  $\lambda, \mu, \sigma$ , and  $\omega$  are some real constants. We can therefore understand the possible qualitative behaviors of a planar linear system by sketching the different possible phase portraits corresponding to each of these matrices.

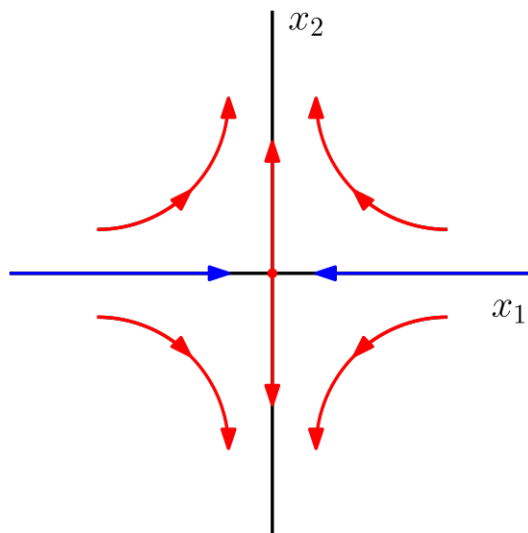
In what follows, we'll develop broad categories into which the different possible phase portraits fall. In doing this, we'll have *classified* all qualitative behaviors (up to change of coordinates) of a planar LTI system. Along the way, we'll keep a special eye out for commonalities between the different phase portraits—this will ultimately lead us to the invariant subspace theorem, one of the main results of this section. Note that here, although we sketch all phase portraits for continuous-time systems, analogous portraits can be drawn for discrete-time systems.

### 3.2.2.1 Saddle

First, we'll examine the matrix,

$$\begin{bmatrix} \lambda & 0 \\ 0 & \mu \end{bmatrix}, \quad (3.113)$$

in the case where we have one stable eigenvalue,  $\lambda < 0$  and one unstable eigenvalue,  $\mu > 0$ . Plotting out trajectories for a few initial conditions yields the phase portrait in Figure 3.3. What's going on in this portrait? Here, we have one “stable mode,” sketched in blue



**Fig. 3.3** The phase portrait corresponding to a diagonal matrix where one eigenvalue is stable and the other is unstable. Such a portrait is called a *saddle*.

(corresponding to  $\lambda$ ), and one “unstable mode,” sketched in red (corresponding to  $\mu$ ). As we can see from the portrait, any trajectory starting along the stable mode converges to the

origin, while any trajectory starting along the unstable mode diverge to  $\infty$ . Any trajectory not starting on the two axes is eventually “pulled away” by the unstable mode to infinity. Since the trajectories along the  $x_1$ -axis all converge and the trajectories along the  $x_2$ -axis all diverge, this system is said to have a *saddle* at the origin.

Geometrically, what can we say about this system? We know that the matrix  $A$  has two linearly independent eigenvectors—one along the  $x_1$ -axis and the other along the  $x_2$ -axis. We note that any trajectory starting along the  $x_1$ -axis will remain on the  $x_1$ -axis for all time. Likewise, we see the same behavior for  $x_2$ . Thus, it seems like the eigenvectors of the  $A$  matrix exhibit some sort of *invariant* behavior with respect to the state transition map—if we start on one of these axes, we never leave! Further, all “straight lines” in the phase portrait correspond to eigenvectors of  $A$ . Let’s keep an eye out for these behaviors as we continue.

### 3.2.2.2 Node

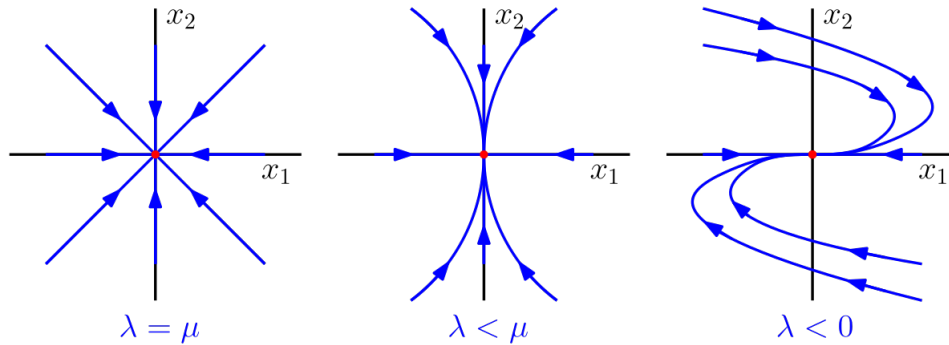
Above, we considered the case where we had one stable mode and one unstable mode—this led to a phase portrait in which we had a *saddle* at the origin. Now, we’ll consider the case where both  $\lambda$  and  $\mu$  are stable. In particular, we’ll examine the cases,

$$\text{Case 1: } \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}, \lambda = \mu < 0 \quad (3.114)$$

$$\text{Case 2: } \begin{bmatrix} \lambda & 0 \\ 0 & \mu \end{bmatrix}, \lambda < \mu < 0 \quad (3.115)$$

$$\text{Case 3: } \begin{bmatrix} \lambda & 1 \\ 0 & \lambda \end{bmatrix}, \lambda < 0. \quad (3.116)$$

The phase portraits corresponding to each of these cases are plotted from left to right in Figure 3.4. Since each of the three cases corresponds to a *stable* origin, the origin is said to



**Fig. 3.4** Three *stable nodes*. In the first case, one has a stable, diagonal matrix with a repeated eigenvalue, in the second case a stable, diagonal matrix where one eigenvalue dominates, and in the third a stable Jordan block.

be a *stable node* of each of the portraits.

Let's examine the three portraits. In the first portrait, where  $\lambda = \mu$  and both  $\lambda, \mu < 0$ , trajectories approach the origin at a uniform rate and come in along straight lines. Here, if we start along any ray leaving the origin, we'll remain along that ray all the time. How can we connect this to the eigenvector observations we made above? Since  $A = \lambda I_2$ , any nonzero vector is an eigenvector of  $A$ ! Therefore, we *again* observe that once we start along an eigenvector, we'll remain on that eigenvector for all time, and that all straight lines in the portrait correspond to eigenvectors.

In the second portrait, when  $\lambda < \mu < 0$ , the  $\lambda$  mode pulls trajectories towards the origin faster than the  $\mu$  mode. This results in the phase portrait becoming "warped," with trajectories bending towards the  $x_2$  axis due to the dominance of  $\lambda$  over  $\mu$ . Here, we again notice the eigenvector behavior—starting along the  $x_1$  or  $x_2$  axis means we will stay there for all time.

In the third portrait, we have a Jordan block. As such,  $A$  only has a single eigenvector along the  $x_1$ -axis. Trajectories starting on this eigenvector remain on the eigenvector, and all other trajectories converge to the origin.

In each of these three cases, a little bit of experimentation reveals the following: the magnitude of  $\lambda$  controls the rate at which trajectories converge to the origin— $\lambda$  with a higher magnitude results in faster convergence. This is consistent with the analysis we performed in the previous section. These three cases are all examples of portraits with *stable nodes*. If instead,  $\lambda$  and  $\mu$  are both picked to be greater than zero, the origin is called an *unstable node* of the portrait. In this case, the arrows of each phase portrait are reversed.

### 3.2.2.3 Focus

Next, we consider the matrix

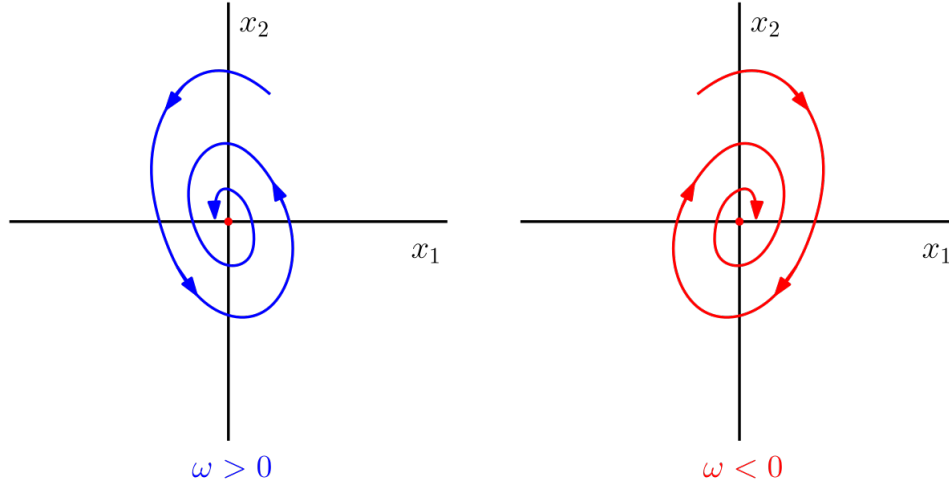
$$A = \begin{bmatrix} \sigma & -\omega \\ \omega & \sigma \end{bmatrix}, \quad (3.117)$$

which has complex eigenvalues  $\sigma \pm j\omega$ . We plot two phase portraits for this system: one in which  $\sigma < 0, \omega > 0$  and one in which  $\sigma < 0, \omega < 0$ . In both cases, since the real part of the eigenvalues of  $A$  is less than zero, the origin is an exponentially stable equilibrium point. But, the presence of a nonzero complex term,  $\omega$ , means that trajectories don't converge to the origin along straight lines, but rather spiral in. This is consistent with our computation of the exponential of a complex Jordan block. Evaluating the exponential of  $A$ , one has

$$\exp(At) = e^{\sigma t} \begin{bmatrix} \cos(\omega t) & -\sin(\omega t) \\ \sin(\omega t) & \cos(\omega t) \end{bmatrix}. \quad (3.118)$$

Thus, we observe that  $\sigma$  controls how quickly the spiral converges to the origin, while  $\omega$  controls the rate of spin around the origin (higher  $\omega$  will lead to faster rotation). Examining the portraits, we note that the sign of  $\omega$  determines the direction of rotation. For  $\omega > 0$ , the trajectories spiral in a counterclockwise direction, while for  $\omega < 0$ , the trajectories spiral in a clockwise direction.

In this case, where we have trajectories spiraling in towards the origin, we refer to the origin as a *stable focus* of the phase portrait. If instead we had  $\sigma > 0$ , the trajectories would spiral away from the origin, and we would refer to the origin as an *unstable focus*.



**Fig. 3.5** Two *stable foci*. Since  $\sigma < 0$  and  $\omega \neq 0$ , trajectories of the system “spiral in” towards the origin. The direction of the spiral (clockwise versus counterclockwise) is determined by the sign of  $\omega$ .

#### 3.2.2.4 Center

Next, we consider the example,

$$A = \begin{bmatrix} 0 & -\omega \\ \omega & 0 \end{bmatrix}, \quad (3.119)$$

of a  $2 \times 2$  matrix with purely imaginary eigenvalues  $\pm j\omega$ . The possible phase portraits for this system with nonzero  $\omega$  are plotted in Figure 3.6. In this case, each phase portrait is composed of concentric circles, centered at the origin, all of which rotate in the same direction. As such, the phase portrait is said to have a *center at the origin*.

We notice that for  $\omega > 0$ , the rotation is counterclockwise, while for  $\omega < 0$ , the rotation is clockwise. This is consistent with our observations from the case of a stable focus, and can also be confirmed by directly computing the exponential of  $A$ . As in the case of the focus, the rate of rotation is controlled by the magnitude of  $\omega$ .

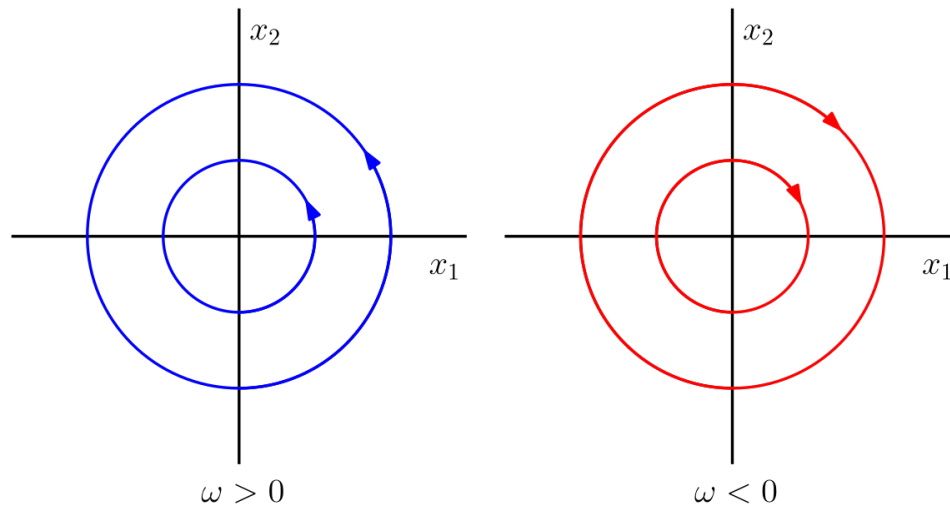
#### 3.2.2.5 Degenerate Equilibria

In the four examples above, we exclusively considered cases in which the origin,  $x_e = 0$ , is the only equilibrium point of the system. What happens if there are other equilibria? In the case where the system has multiple equilibria, the origin is referred to as the *degenerate equilibrium*. By studying the possible phase portraits associated with the matrices,

$$\begin{bmatrix} \lambda & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad (3.120)$$

all of which have equilibrium points other than  $x_e = 0$ , one can complete the job of classifying the possible phase portraits of a planar, LTI system—the details of this task are requested





**Fig. 3.6** Two *centers at the origin*. When  $A$  has purely imaginary eigenvalues, the portraits are composed of concentric circles, centered at the origin. The direction of rotation is determined by the sign of  $\omega$ —positive  $\omega$  means counterclockwise rotation, while negative  $\omega$  means clockwise rotation.

in Problem 3.12. This case is made particularly interesting by the fact that there is not one but rather an *infinite* number of equilibria.

### 3.2.2.6 Phase Portrait Takeaways

Let's summarize what we've learned from our analysis of planar phase portraits. First, we found that the real parts of the eigenvalues of  $A$  control the rate of decay to the origin, while the complex parts of the eigenvalues of  $A$  control the rate of rotation about the origin. Secondly, we seemed to find that trajectories *starting* along an eigenvector would remain along that eigenvector for all time.

It's this second observation—that the span of an eigenvector seems to be *invariant* under the evolution of the system—will lead us to an important theorem called the *invariant subspace theorem*. This result, which is the subject of the next subsection, lets us decompose the state space of *any* LTI system into invariant subspaces that are determined by the generalized eigenspaces (and therefore generalized eigenvectors) of  $A$ .

### 3.2.3 The Invariant Subspace Theorem

Now, we return to the general, non-planar case. Notably, in the various phase portraits we examined above, we found that any trajectory which *starts* on an eigenvector of  $A$  will *remain* on that eigenvector for all time. How does this observation generalize to the case of a linear system on  $\mathbb{R}^n$ ? In order to precisely answer this question, we first need to introduce a rigorous notion of *invariance* for dynamical systems.

**Definition 3.8 (Invariant Set)** Consider an unforced dynamical system with state transition map  $\varphi : \mathcal{T} \times \mathcal{T} \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ . A set  $\mathcal{S} \subseteq \mathbb{R}^n$  is said to be:

1. Invariant with respect to  $\varphi$  if, for all  $t_0 \in \mathcal{T}$ ,

$$x_0 \in \mathcal{S} \implies \varphi(t, t_0, x_0) \in \mathcal{S}, \forall t \in \mathcal{T}. \quad (3.121)$$

2. Forward invariant with respect to  $\varphi$  if, for all  $t_0 \in \mathcal{T}$ ,

$$x_0 \in \mathcal{S} \implies \varphi(t, t_0, x_0) \in \mathcal{S}, \forall t \geq t_0 \in \mathcal{T}. \quad (3.122)$$

*Remark 3.15* Note that the definition of *invariance*, as posed above, requires a state transition map which is defined on all of  $\mathcal{T} \times \mathcal{T} \times \mathbb{R}^n$ . Recall that, in general, the state transition is only defined on  $\mathbf{T} \times \mathbb{R}^n$  (where  $\mathbf{T} = \{(t_1, t_0) \in \mathcal{T} \times \mathcal{T} : t_1 \geq t_0\}$ ). For continuous-time systems, since unique solutions will exist for all  $(t_1, t_0, x_0) \in \mathcal{T} \times \mathcal{T} \times \mathbb{R}^n$ , invariance is always a well-defined concept. For discrete-time systems, however, we know this is not the case without additional assumptions on the invertibility of the system data. However, we can always reason about *forward invariance* for any state transition map  $\varphi$  on  $\mathbf{T} \times \mathbb{R}^n$ , since forward invariance only requires us to examine pairs of times  $t \geq t_0$ .

Thus, a set is *invariant* with respect to  $\varphi$  if the initial condition belonging to the set implies that the trajectory belongs to the set for *all time*. Here, it's important to note that “all time” means *all of*  $\mathcal{T}$ , not just  $t \geq t_0$ —this case is referred to as *forward invariance*.

In the planar case, it seemed as if the span of each eigenvector of  $A$  was invariant with respect to the state transition map  $\varphi$  of  $\dot{x} = Ax$ . If one started along an eigenvector corresponding to a *stable* eigenvalue, for instance, one would stay along that eigenvector for all time. Likewise, starting along an unstable eigenvector means staying along the unstable eigenvector for all time. In order to generalize from  $\mathbb{R}^2$  to  $\mathbb{R}^n$ , we move away from single stable and unstable *vectors* towards stable and unstable *subspaces*.

**Definition 3.9 (Stable, Unstable, and Center Subspaces)** Consider a linear, time-invariant system with state equation  $\dot{x} = Ax$ ,  $A \in \mathbb{R}^{n \times n}$ . Suppose  $A$  has  $p$  distinct real eigenvalues and  $k$  conjugate pairs of distinct complex eigenvalues,

$$\underbrace{\{\lambda_1, \dots, \lambda_p\}}_{\text{Real}} \cup \underbrace{\{\lambda_{p+1}, \bar{\lambda}_{p+1}, \dots, \lambda_{p+k}, \bar{\lambda}_{p+k}\}}_{\text{Complex}} \subseteq \mathbb{C}. \quad (3.123)$$

Let  $\mathcal{B} = \{u_1, \dots, u_m, u_{m+1}, v_{m+1}, \dots, u_{m+q}, v_{m+q}\}$  be a basis for  $\mathbb{R}^n$  which takes  $A$  into a real Jordan form. The stable, unstable, and center subspaces with respect to  $A$  are defined,

1. Stable subspace:  $E^s = \text{span}\{u_i, v_i : \text{Re}(\lambda_i) < 0\} \subseteq \mathbb{R}^n$ .
2. Unstable subspace:  $E^u = \text{span}\{u_i, v_i : \text{Re}(\lambda_i) > 0\} \subseteq \mathbb{R}^n$ .
3. Center Subspace:  $E^c = \text{span}\{u_i, v_i : \text{Re}(\lambda_i) = 0\} \subseteq \mathbb{R}^n$ .

*Remark 3.16* At least superficially—it appears as if the definitions of the stable, unstable, and center subspaces depend on the choice of basis which transforms  $A$  into Jordan form! For instance, if one defines  $K^- = K_{\lambda_1}(A) \oplus \dots \oplus K_{\lambda_m}(A) : \text{Re}(\lambda_i) < 0$ , the direct sum of all generalized eigenspaces corresponding to stable eigenvalues (see the definition of a direct sum below), then the stable subspace is equal to  $K^- \cap \mathbb{R}^n$ . Similar constructions can be made for the unstable and center subspaces. Thus, the stable, unstable, and center subspaces are intrinsic to the matrix  $A$ , and not to the choice of basis  $\mathcal{B}$  used in Definition

**3.9.** However, Definition 3.9 is often more convenient to work with than the equivalent generalized eigenspace definition, since it is somewhat more explicit.

In order to generalize this result to  $\mathbb{R}^n$ , we'd like to show that we can decompose  $\mathbb{R}^n$  into a sum of spaces, each of which is invariant with respect to  $\varphi$ . Before we prove this, however, we first must introduce a precise notion of *decomposition*.

**Definition 3.10 (Direct Sum of Subspaces)** Let  $V$  be a vector space and  $V_1, \dots, V_m$  be subspaces of  $V$ .  $V$  is said to be the *sum* of  $V_1, \dots, V_m$ , denoted

$$V = V_1 + \dots + V_m, \quad (3.124)$$

if for all  $v \in V$ , there exist  $v_i \in V_i$ ,  $i = 1, \dots, m$  for which  $v = v_1 + \dots + v_m$ . If for each  $v \in V$ , there exist unique  $v_i \in V_i$ ,  $i = 1, \dots, m$  for which  $v = v_1 + \dots + v_m$ , then  $V$  is said to be the *direct sum* of  $V_1, \dots, V_m$ , denoted

$$V = V_1 \oplus \dots \oplus V_m. \quad (3.125)$$

Using the concept of a direct sum, we can precisely state what it means for a collection of subspaces to form a decomposition of a larger space. With this in mind, we state and prove the invariant subspace theorem.

**Theorem 3.7 (Invariant Subspace Theorem)** Consider a linear, time-invariant system with state equation  $\dot{x} = Ax$ ,  $A \in \mathbb{R}^{n \times n}$ , where  $A$  has stable, unstable, and center subspaces  $E^s, E^u, E^c$ . Then,  $\mathbb{R}^n$  may be decomposed,

$$\mathbb{R}^n = E^s \oplus E^u \oplus E^c. \quad (3.126)$$

Further, each of  $E^s, E^u, E^c$  is invariant with respect to  $\dot{x} = Ax$ .

**Proof** First, we show that  $\mathbb{R}^n = E^s \oplus E^u \oplus E^c$ . Recall that the set of vectors,

$$\mathcal{B} = \{u_1, \dots, u_m, u_{m+1}, v_{m+1}, \dots, u_{m+q}, v_{m+q}\}, \quad (3.127)$$

which defines each of  $E^s, E^u, E^c$ , forms a basis for  $\mathbb{R}^n$  in which  $[A]_{\mathcal{B}}$  is in real Jordan form. Since  $\mathcal{B}$  forms a basis, for all  $v \in \mathbb{R}^n$ , there exist unique constants  $c_1^u, \dots, c_m^u, c_{m+1}^u, c_{m+1}^v, \dots, c_{m+q}^u, c_{m+q}^v \in \mathbb{R}$  for which

$$v = \sum_{i=1}^{m+q} c_i^u u_i + \sum_{i=1}^q c_{m+i}^v v_i. \quad (3.128)$$

Reorganizing this sum by grouping together the  $u_i, v_i$  terms belonging to each of the stable, unstable, and center subspaces, it follows that there exist unique  $w^s \in E^s, w^u \in E^u, w^c \in E^c$  for which  $v = w^s + w^u + w^c$ . It follows that  $\mathbb{R}^n = E^s \oplus E^u \oplus E^c$ .

Next, we prove that each subspace is invariant with respect to the transition map of  $\dot{x} = Ax$ . Since this system is time-invariant, it's sufficient to show invariance under  $\exp(At)$  for all  $t \in \mathbb{R}$ —there's no need to verify invariance for any other  $t_0 \in \mathbb{R}$ . Without loss of generality, assume that the real Jordan form of  $A$  is organized such that,

$$J = \text{blkdiag}(J^s, J^u, J^c), \quad (3.129)$$

where  $J^s$  contains all Jordan blocks corresponding to stable eigenvalues,  $J^u$  contains all Jordan blocks with unstable eigenvalues, and  $J^c$  all Jordan blocks with center eigenvalues (given any real Jordan form, we can always rearrange the blocks and the order of the basis vectors such that this is true).

Let  $T \in \mathbb{R}^{n \times n}$  be the matrix for which  $J = TAT^{-1}$ . Due to the ordering of the Jordan blocks, any vector  $v \in E^s$  gets transformed to a block vector of the form  $Tv = (v^s, 0, 0)$ . Likewise, any  $v \in E^u$  gets transformed to  $Tv = (0, v^u, 0)$ , and  $v \in E^c$  to  $Tv = (0, 0, v^c)$ . Let  $v \in E^s$ . Then, it follows that

$$v \in E^s \Rightarrow \exp(At)v = T^{-1} \exp(Jt)Tv = T^{-1}(\exp(J^s t)v^s, 0, 0) \quad (3.130)$$

$$v \in E^u \Rightarrow \exp(At)v = T^{-1} \exp(Jt)Tv = T^{-1}(0, \exp(J^u t)v^u, 0) \quad (3.131)$$

$$v \in E^c \Rightarrow \exp(At)v = T^{-1} \exp(Jt)Tv = T^{-1}(0, 0, \exp(J^c t)v^c). \quad (3.132)$$

This implies that, for all  $t \in \mathbb{R}$ ,

$$v \in E^s \Rightarrow T \exp(At)v = (\exp(J^s t)v^s, 0, 0) \Rightarrow \exp(At)v \in E^s \quad (3.133)$$

$$v \in E^u \Rightarrow T \exp(At)v = (0, \exp(J^u t)v^u, 0) \Rightarrow \exp(At)v \in E^u \quad (3.134)$$

$$v \in E^c \Rightarrow T \exp(At)v = (0, 0, \exp(J^c t)v^c) \Rightarrow \exp(At)v \in E^c. \quad (3.135)$$

We conclude that each subspace must be invariant under multiplication by  $\exp(At)$ . Since  $\varphi(t, 0, x_0) = \exp(At)x_0$ , it follows that each of  $E^s, E^u, E^c$  is invariant with respect to  $\varphi$ .  $\square$

*Remark 3.17* The proof of the invariant subspace theorem presented above is not the only one possible! One can also prove the invariant subspace theorem by showing each generalized eigenspace of  $A$  is invariant under  $\exp(At)$  for all  $t \in \mathbb{R}$ , and arguing that each of  $E^s, E^u, E^c$  is invariant as a result. The details of this proof are requested in Problem 3.7.

The invariant subspace theorem therefore lets us understand how trajectories of a system will behave depending on where they start. If a trajectory starts in the stable subspace, for instance, it must converge exponentially to the origin. On the other hand, if a trajectory starts in the unstable subspace, it will diverge exponentially away from the origin. For a continuous-time, LTI system, we can rephrase this by saying that the trajectory converges to zero as  $t \rightarrow -\infty$ . We formalize these ideas in the following result.

**Corollary 3.2 (Convergence & Invariant Subspace)** *Consider a linear, time-invariant system  $\dot{x} = Ax$ ,  $x \in \mathbb{R}^n$ .*

1. Convergence: If  $x_0 \in E^s$ , then  $\lim_{t \rightarrow \infty} \varphi(t, 0, x_0) = 0$ .
2. Divergence: If  $x_0 \in E^u$ , then  $\lim_{t \rightarrow -\infty} \varphi(t, 0, x_0) = 0$ .

**Exercise 3.16** Prove Corollary 3.2.

What other some other important consequences of the invariant subspace theorem? Based on the direct sum decomposition of  $\mathbb{R}^n$ , we can decompose individual trajectories into stable, unstable, and center components.

**Corollary 3.3 (Trajectory Decomposition)** *Consider a linear, time-invariant system  $\dot{x} = Ax$ ,  $x \in \mathbb{R}^n$ . For all  $x_0 \in \mathbb{R}^n$ , there exist unique  $x_0^s \in E^s$ ,  $x_0^u \in E^u$ ,  $x_0^c \in E^c$  for which,*

$$\varphi(t, 0, x_0) = \varphi(t, 0, x_0^s) + \varphi(t, 0, x_0^u) + \varphi(t, 0, x_0^c), \quad \forall t \in \mathbb{R}. \quad (3.136)$$

*Thus, every trajectory can be decomposed into stable, unstable, and center components.*

**Proof** Fix  $x_0 \in \mathbb{R}^n$ . Since  $\mathbb{R}^n = E^s \oplus E^u \oplus E^c$ , there exist unique  $x_0^s, x_0^u, x_0^c$  for which  $x_0 = x_0^s + x_0^u + x_0^c$ . By linearity in the initial condition, one has

$$\varphi(t, 0, x_0) = \varphi(t, 0, x_0^s) + \varphi(t, 0, x_0^u) + \varphi(t, 0, x_0^c), \quad \forall t \in \mathbb{R}. \quad (3.137)$$

Since each subspace is invariant, it follows that  $\varphi(t, 0, x_0^s) \in E^s$ ,  $\varphi(t, 0, x_0^u) \in E^u$ , and  $\varphi(t, 0, x_0^c) \in E^c$  for all  $t \in \mathbb{R}$ .  $\square$

As another consequence of the invariant subspace theorem, we find another, equivalent way of characterizing the global exponential stability of the equilibrium  $x_e = 0$ .

**Corollary 3.4 (Subspace Characterization of Stability)** *Consider a linear, time-invariant system  $\dot{x} = Ax$ ,  $x \in \mathbb{R}^n$ . The equilibrium  $x_e = 0$  of  $\dot{x} = Ax$  is globally exponentially stable if and only if  $E^s = \mathbb{R}^n$ .*

**Exercise 3.17** Prove Corollary 3.4.

Let's quickly summarize what the invariant subspace approach to stability tells us. The invariant subspace theorem tells us that we can decompose the state space of a linear, time-invariant system into invariant subspaces, each of which describes a different mode of stability of the system. If the initial condition of a system falls exclusively into one of these subspaces, we can exactly understand which mode of stability the trajectory will have. For a general initial condition, we can uniquely decompose the solution into components contained in each subspace.

It's important to note that, although we've proved all of our results in continuous time, the proofs of each result relied solely on the properties of the Jordan form. Thus, one can easily adapt the theory of invariant subspaces from continuous-time to discrete-time, simply by changing the eigenvalue conditions defining each subspace and changing invariance to forward invariance. The details of this are requested in Problem 3.13.

### 3.2.4 The Lyapunov Equation

Above, we showed that we can use *invariant subspaces* to develop another, equivalent characterization of stability through the invariant subspace theorem. In this subsection, we take advantage of a different type of invariance to study the stability of linear systems through the method of *Lyapunov functions*. The main idea behind this method is the following: by defining a special function on the state space, whose sublevel sets are forward invariant, we can establish the Lyapunov, asymptotic, and exponential stability of equilibria.

In order to motivate this technique, we return to the definition of Lyapunov stability. Recall that an equilibrium point  $x_e = 0$  is Lyapunov stable for an unforced system with state transition map  $\varphi$  if, for each fixed  $\epsilon > 0$ , there exists a  $\delta > 0$  such that *starting* within distance  $\delta$  of 0 means the system will *stay* within distance  $\epsilon$  of 0 for all time. That is,

$$\|x_0\| < \delta \implies \|\varphi(t, t_0, x_0)\| < \epsilon, \quad \forall t \geq t_0. \quad (3.138)$$

What does this have to do with *invariance*? Taking a closer look at the definition, we can rewrite the condition as,

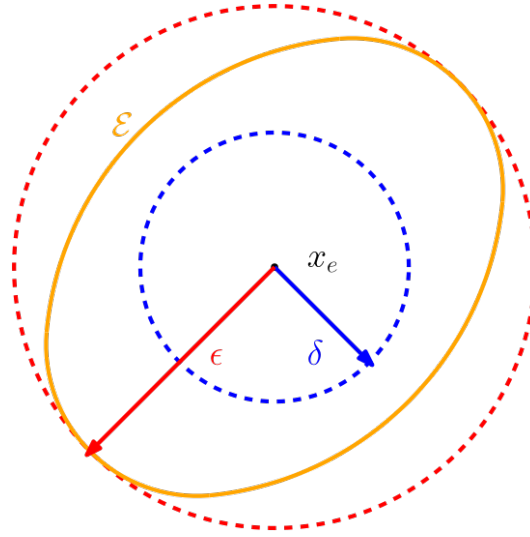
$$x_0 \in B_\delta(0) \implies \varphi(t, t_0, x_0) \in B_\epsilon(0), \quad \forall t \geq t_0. \quad (3.139)$$

Thus, the condition  $\varphi(t, t_0, x_0) \in B_\epsilon(0) \forall t \geq t_0$  seems somewhat related to that of invariance! Let's work on simplifying this statement even further. For an unforced linear, time-invariant system with state transition matrix  $\Phi$ , we know that  $\varphi(t, t_0, x_0) = \Phi(t, t_0)x_0$ . Thus, the condition  $x_0 \in B_\delta(0) \Rightarrow \varphi(t, t_0, x_0) \in B_\epsilon(0)$  is equivalently written,

$$\Phi(t, t_0) \cdot B_\delta(0) \subseteq B_\epsilon(0), \forall t \geq t_0, \quad (3.140)$$

where  $\Phi(t, t_0) \cdot B_\delta(0)$  refers to the set  $\{\Phi(t, t_0)v : v \in B_\delta(0)\}$ . What is the shape of the set  $\Phi(t, t_0) \cdot B_\delta(0)$ ? As a consequence of the singular value decomposition, we know that, for any matrix  $C \in \mathbb{R}^{n \times n}$ , the set  $C \cdot B_\delta(0)$  has the shape of an *ellipsoid*, the higher-dimensional analogue of an ellipsoid. Applying this fact to the problem at hand, we conclude that, for all  $t \geq t_0$ , the set  $\Phi(t, t_0) \cdot B_\delta(0)$  must have the shape of an ellipsoid!

This leads us to the following idea. Suppose that, for any  $\epsilon > 0$ , we can find a forward invariant ellipsoid,  $\mathcal{E} \subseteq B_\epsilon(0)$ , i.e. an ellipsoid for which  $x_0 \in \mathcal{E}$  implies  $\varphi(t, t_0, x_0) \in \mathcal{E}$  for all  $t \geq t_0$ . Then, by picking any  $\delta > 0$  such that  $B_\delta(0) \subseteq \mathcal{E}$ , we will meet the requirement of  $x_0 \in B_\delta(0) \Rightarrow \varphi(t, t_0, x_0) \in B_\epsilon(0)$ . Thus, using this ellipsoid method, it *seems* as if we should be able to establish the Lyapunov stability of  $x_e = 0$ !



**Fig. 3.7** Above, we inscribe an ellipsoid  $\mathcal{E}$  into the  $\epsilon$ -ball  $B_\epsilon(0)$ . If we're guaranteed that starting in  $\mathcal{E}$  implies we will *stay* in  $\mathcal{E}$  for all  $t \geq t_0$ , then any  $\delta$  such that  $B_\delta(0) \subseteq \mathcal{E}$  will meet the conditions for Lyapunov stability!

Let's organize the pieces we need to formalize this reasoning. First, we need a method of finding ellipsoids that are contained within any given  $\epsilon$ -ball. Secondly, we need a method of ensuring these ellipsoids are *forward invariant*—i.e. that  $x_0 \in \mathcal{E}$  implies  $\varphi(t, t_0, x_0) \in \mathcal{E}$  for all  $t \geq t_0$ . Then, we need to ensure that we can pick a  $\delta$  ball inside any such ellipsoid. Once we have these three components, we *should* be able to show that this “ellipsoid” method gives us a way to prove stability.

With this said, we begin the process of formalizing this “ellipsoid” approach to stability. First, we recall the formal definition of an ellipsoid.

**Definition 3.11 (Ellipsoid)** An ellipsoid in  $\mathbb{R}^n$  is a set of the form

$$\mathcal{E} = \{x \in \mathbb{R}^n : x^\top P x \leq c\}, \quad (3.141)$$

where  $c > 0$  is a fixed positive scalar and  $P \in \mathbb{S}^n, P \succ 0$  is a fixed positive definite matrix.

*Remark 3.18* Notice that we assume all ellipsoids to be *centered at the origin*.

*Remark 3.19*  $\mathbb{S}^n$  refers to the set of symmetric,  $n \times n$  matrices with real entries.

This definition makes apparent an easy way of generating a family of ellipsoids that fit within  $\epsilon$ -balls centered at the origin. For a fixed positive definite matrix  $P \in \mathbb{S}^n, P \succ 0$ , consider the family of ellipsoids  $\mathcal{E}_c = \{x : x^\top P x \leq c\}$ . One may show that, for any  $\epsilon > 0$ , there exists a  $c > 0$  for which  $\mathcal{E}_c \subseteq B_\epsilon(0)$ .

**Exercise 3.18** Verify that for a family of ellipsoids,  $\mathcal{E}_c = \{x : x^\top P x \leq c\}$ , for any  $\epsilon > 0$  one can pick a  $c > 0$  such that  $\mathcal{E}_c \subseteq B_\epsilon(0)$ .

This finishes our first item, regarding generation of ellipsoids that fit within arbitrarily small  $\epsilon$ -balls. Now, we move on to the second item: finding conditions under which each ellipsoid is forward invariant under the dynamics of the system. How might we do this? Taking another look at the definition of the ellipsoid,

$$\mathcal{E}_c = \{x \in \mathbb{R}^n : x^\top P x \leq c\}, \quad (3.142)$$

it *seems* like  $\mathcal{E}_c$  can be written in terms of a function! Define a function  $V : \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $V(x) = x^\top P x$ . Based on the above, we find that  $\mathcal{E}_c$  is equivalently described,

$$\mathcal{E}_c = \{x \in \mathbb{R}^n : V(x) \leq c\}. \quad (3.143)$$

Thus, in order for  $\mathcal{E}_c$  to be forward invariant with respect to the state transition map, it's sufficient to have  $V(\varphi(t, t_0, x_0)) \leq V(x_0)$  for all  $t \geq t_0$  and all  $x_0 \in \mathbb{R}^n$ . If this is the case,  $x_0 \in \mathcal{E}_c$  implies  $V(\varphi(t, t_0, x_0)) \leq V(x_0) \leq c$ , which in turn implies  $\varphi(t, t_0, x_0) \in \mathcal{E}_c$  for all  $t \geq t_0$ . This observation leads us to the following definition.

**Definition 3.12 (Quadratic Lyapunov Function)** Consider an unforced system with state transition map  $\varphi : \mathbf{T} \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ . A function  $V : \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $V(x) = x^\top P x$ , where  $P \in \mathbb{S}^n, P \succ 0$ , is said to be a quadratic Lyapunov function for the system if for all  $t_0 \in \mathcal{T}$  and  $x_0 \in \mathbb{R}^n$ ,

$$V(\varphi(t, t_0, x_0)) \leq V(x_0), \quad \forall t \geq t_0. \quad (3.144)$$

Thus, a given quadratic function  $V$  is said to be a *quadratic Lyapunov function* for the system if  $V$  doesn't increase along trajectories of the system. Now, let's put this all together to show we can use quadratic Lyapunov functions to verify the stability of the zero equilibrium.

**Theorem 3.8 (Quadratic Lyapunov Theorem)** Consider an unforced system with state transition map  $\varphi : \mathbf{T} \times \mathbb{R}^n \rightarrow \mathbb{R}^n$  and equilibrium  $x_e = 0$ . If there exists a quadratic Lyapunov function for the system, then the equilibrium  $x_e = 0$  is Lyapunov stable.

*Remark 3.20* This theorem is easily generalized to the case where  $x_e \neq 0$  by performing a change of coordinates  $z = x - x_e$ .

**Proof** Suppose there exists a quadratic Lyapunov function,  $V(x) = x^\top Px$ , for the system. Fix  $\epsilon > 0$ , and choose an ellipsoid  $\mathcal{E}_\epsilon$  for which  $\mathcal{E}_\epsilon \subseteq B_\epsilon(0)$ . Since  $V$  is a quadratic Lyapunov function,  $\mathcal{E}_\epsilon$  must be forward invariant with respect to  $\varphi$ . Now, choose  $\delta$  such that  $B_\delta(0) \subseteq \mathcal{E}_\epsilon$  (such a  $\delta$  is guaranteed to exist since the ellipse  $\mathcal{E}_\epsilon$  is centered at zero and has nonempty interior). Since  $\mathcal{E}_\epsilon$  is forward invariant, for any  $x_0 \in B_\delta(0) \subseteq \mathcal{E}_\epsilon$  and  $t_0 \in \mathcal{T}$ , it follows that  $\varphi(t, t_0, x_0) \in \mathcal{E}_\epsilon \subseteq B_\epsilon(0)$  for all  $t \geq t_0$ . Thus, the equilibrium  $x_e = 0$  is Lyapunov stable.  $\square$

Thus, we observe that the existence of a quadratic Lyapunov function implies the stability of the equilibrium  $x_e = 0$ . So far, however, this theorem doesn't buy us much insight into stability—it still involves an inequality using the entire state transition map,  $\varphi$ , which might be challenging to verify. How can we eliminate the state transition map from the statement of this theorem?

Let's perform a little bit of analysis in the linear, time-invariant case to see if we can simplify the inequality  $V(\varphi(t, t_0, x_0)) \leq V(x_0)$ . In the continuous-time, LTI case, where  $\dot{x} = Ax$ , we can guarantee that this inequality holds if the *derivative* of  $V$  along trajectories of the system is less than or equal to zero. That is, it's sufficient to have,

$$\dot{V}(x(t)) = \frac{d}{dt}(x(t)^\top Px(t)) = x^\top(t)A^\top Px(t) + x^\top(t)PAx(t) \leq 0 \quad (3.145)$$

$$x^\top(t)(A^\top P + PA)x^\top(t) \leq 0. \quad (3.146)$$

Likewise, in the discrete-time case, for  $x[k+1] = Ax[k]$ , we can guarantee the inequality holds if at each step,  $V$  doesn't increase. That is, it's sufficient to have,

$$V[k+1] = x^\top[k+1]Px[k+1] = x^\top[k]A^\top PAx[k] \leq x^\top[k]Px[k] \quad (3.147)$$

$$x^\top[k](A^\top PA - P)x^\top[k] \leq 0. \quad (3.148)$$

Interestingly, we find that, when we evaluate the decrease conditions, we get *algebraic* expressions involving  $A$  and  $P$ . Since  $x^\top(t)(A^\top P + PA)x^\top(t) \leq 0$  if  $A^\top P + PA \preceq 0$ , and  $x^\top[k](A^\top PA - P)x^\top[k] \leq 0$  if  $A^\top PA - P \preceq 0$ , we conjecture,

$$\exists P \succ 0 \text{ s.t. } A^\top P + PA \preceq 0 \implies x_e = 0 \text{ is SISL for } \dot{x} = Ax \quad (3.149)$$

$$\exists P \succ 0 \text{ s.t. } A^\top PA - P \preceq 0 \implies x_e = 0 \text{ is SISL for } x[k+1] = Ax[k]. \quad (3.150)$$

Unlike the quadratic Lyapunov theorem we stated above, the equations posed in this conjecture are *entirely* algebraic! We claim that *all* we need to do to prove stability is establish there exists a positive definite  $P$  satisfying  $A^\top P + PA \preceq 0$  (in the continuous-time case) or  $A^\top PA - P \preceq 0$  (in the discrete-time case).

If we claim that the non-strict inequalities  $A^\top P + PA \preceq 0$  and  $A^\top PA - P \preceq 0$  imply Lyapunov stability, what can we claim about *stronger* modes of stability? A reasonable guess is that, if a quadratic Lyapunov function *strictly decreases* along non-equilibrium trajectories, then we get *asymptotic* stability. This is a reasonable guess, since for  $V(x) = x^\top Px$ ,  $\dot{V}(x(t)) < 0$  for  $x(t) \neq 0$  seems to imply Lyapunov stability (by our claim above) and that  $x(t) \rightarrow 0$  as  $t \rightarrow \infty$ . Repeating the calculations above with strict inequalities yields,

$$A^\top P + PA \prec 0 \quad (3.151)$$

$$A^\top PA - P \prec 0. \quad (3.152)$$



Using the fact that  $C \prec 0 \Leftrightarrow \exists Q \succ 0$  s.t.  $C \preceq -Q$ , and that  $x_e = 0$  is asymptotically stable for an LTI system if and only if it is exponentially stable, we make the conjectures,

$$\exists P, Q \succ 0 \text{ s.t. } A^\top P + PA \preceq -Q \implies x_e = 0 \text{ is GES for } \dot{x} = Ax \quad (3.153)$$

$$\exists P, Q \succ 0 \text{ s.t. } A^\top PA - P \preceq -Q \implies x_e = 0 \text{ is GES for } x[k+1] = Ax[k]. \quad (3.154)$$

As we'll see shortly, not only are both of these conjectures true, but their converses, among a number of other results, are true as well! Let's begin the process of proving these conjectures and their converses by formally defining each of the expressions we established above.

**Definition 3.13 (Lyapunov Equation/Inequality)** Consider matrices  $A, Q \in \mathbb{R}^{n \times n}$ .

1. The continuous-time Lyapunov equation (CTLE) defined by  $(A, Q)$ , is the equation,

$$A^\top P + PA = -Q, \quad (3.155)$$

in an unknown matrix  $P \in \mathbb{R}^{n \times n}$ . If the equality is traded for an inequality, the expression  $A^\top P + PA \preceq -Q$  is said to be the continuous-time Lyapunov inequality defined by  $(A, Q)$ .

2. The discrete-time Lyapunov equation (DTLE) defined by  $(A, Q)$  is the equation,

$$A^\top PA - P = -Q, \quad (3.156)$$

in an unknown matrix  $P \in \mathbb{R}^{n \times n}$ . If the equality is traded for an inequality, the expression  $A^\top PA - P \preceq -Q$  is said to be the discrete-time Lyapunov inequality defined by  $(A, Q)$ .

*Remark 3.21* The equation  $A^\top P + PA = -Q$  is an instance of a more general matrix equation called a *Sylvester equation*. Likewise, the equation  $A^\top PA - P = -Q$  is an instance of a more general matrix equation called a *Stein equation*.

### 3.2.4.1 The Continuous-Time Lyapunov Equation

Let's begin by analyzing the continuous-time case. In the following result, we show that a matrix  $A \in \mathbb{R}^{n \times n}$  is Hurwitz if and only if there exists a unique solution to the CTLE defined by  $(A, Q)$  for all  $Q \succ 0$ .

**Theorem 3.9 (CTLE Test for Exponential Stability)**  $A \in \mathbb{R}^{n \times n}$  is Hurwitz if and only if, for any  $Q \in \mathbb{S}^n$ ,  $Q \succ 0$ , there exists a unique solution  $P \in \mathbb{S}^n$ ,  $P \succ 0$  to the CTLE,

$$A^\top P + PA = -Q. \quad (3.157)$$

Further, the unique solution to the CTLE is computed,

$$P = \int_0^\infty e^{A^\top t} Q e^{At} dt. \quad (3.158)$$

*Remark 3.22* Recall that a matrix  $A \in \mathbb{R}^{n \times n}$  is said to be *Hurwitz* if all of its eigenvalues have negative, real-components. We proved in the previous section that this is equivalent to global exponential stability of  $x_e = 0$  for  $\dot{x} = Ax$ . Thus, this theorem states that,  $x_e = 0$  is globally exponentially stable if and only if there exists a positive definite solution to the CTLE for all  $Q \succ 0$ .

*Remark 3.23* One can show an even stronger result than that stated above. In particular, if  $A$  is Hurwitz, a unique solution to the CTLE  $(A, Q)$  will exist for *any*  $Q \in \mathbb{R}^{n \times n}$ , not just  $Q \succ 0$ ! We will prove this stronger result to prove one direction of Theorem 3.9.

**Proof** Suppose  $A \in \mathbb{R}^{n \times n}$  is Hurwitz. In order to prove that there exists a unique solution to the CTLE for any  $Q \succ 0$ , we'll first prove a stronger result: that there exists a solution for any  $Q \in \mathbb{R}^{n \times n}$ . Fix a matrix  $Q \in \mathbb{R}^{n \times n}$ . We will show that the given formula for  $P$  provides the unique solution to the CTLE defined by  $(A, Q)$ . First, we note that if  $A$  is Hurwitz, then the integral definition for  $P$  converges, since

$$\|P\| \leq \int_0^\infty \|e^{A^\top t}\| \|Q\| \|e^{At}\| dt < \infty, \quad (3.159)$$

due to the existence of a decaying exponential bound on  $\|\exp(At)\|$ . Now, we prove that  $A^\top P + PA = -Q$ . For the proposed formula for  $P$ , we have,

$$A^\top P + PA = \int_0^\infty A^\top e^{A^\top t} Q e^{At} + e^{A^\top t} Q e^{At} A dt \quad (3.160)$$

$$= \int_0^\infty \frac{d}{dt} [e^{A^\top t} Q e^{At}] dt \quad (3.161)$$

$$= [e^{A^\top t} Q e^{At}]_0^\infty, \quad (3.162)$$

where we recognize the product rule in the first integral and apply fundamental theorem of calculus to calculate the last integral. Since  $A$  is Hurwitz,  $\lim_{t \rightarrow \infty} e^{At} = 0$ . Thus, we have,

$$A^\top P + PA = [e^{A^\top t} Q e^{At}]_0^\infty = 0 - e^{A^\top \cdot 0} Q e^{A \cdot 0} = -IQI = -Q. \quad (3.163)$$

Thus, we conclude that  $P$  solves the CTLE. Now, we show that  $P$  is the *unique* solution. Define a map  $\Pi : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}^{n \times n}$  by,

$$\Pi(X) = A^\top X + XA. \quad (3.164)$$

Above, we showed that for any  $Q \in \mathbb{R}^{n \times n}$ , there exists an  $X$  for which  $\Pi(X) = -Q$ . So, we conclude that  $\Pi$  is a surjective map, with  $\text{range}(\Pi) = \mathbb{R}^{n \times n}$ . Since  $\Pi$  is also linear, we conclude by the fundamental theorem of linear algebra that  $\ker(\Pi) = \{0\}$ . This implies  $\Pi$  is an invertible linear map, which in turn implies  $P$  is the unique solution to the CTLE.

In order to complete the first direction, we must show that, when  $Q \succ 0$ ,  $P \succ 0$  as well. Suppose  $Q \succ 0$ . Then,  $P = \int_0^\infty e^{A^\top t} Q e^{At} dt$  satisfies,

$$x^\top P x = \int_0^\infty x^\top e^{A^\top t} Q e^{At} x dt = \int_0^\infty (e^{At} x)^\top Q (e^{At} x) dt. \quad (3.165)$$

Since  $Q \succ 0$ , the integrand is positive for all nonzero  $x \in \mathbb{R}^n$ , which implies that  $x^\top P x > 0$  for all  $x \neq 0$ . Since  $P$  is also symmetric, we conclude that it must be positive definite. This completes the forward direction.

For the reverse direction, suppose for any given  $Q \succ 0$  there exists a positive definite solution  $P \succ 0$  to the CTLE. Fix a matrix  $Q \succ 0$ , and let  $P \succ 0$  be the corresponding solution of the CTLE. Let  $\lambda$  be an eigenvalue of  $A$  with a (nonzero) eigenvector  $v$ . Then,

$$v^*(A^\top P + PA)v = -v^*Qv \quad (3.166)$$

$$\bar{\lambda}v^*Pv + \lambda v^*Pv = -v^*Qv \quad (3.167)$$

$$2\operatorname{Re}(\lambda)v^*Pv = -v^*Qv \quad (3.168)$$

$$\operatorname{Re}(\lambda) = \frac{-v^*Qv}{2v^*Pv}, \quad (3.169)$$

where we use that  $Av = \lambda v$  implies  $\bar{\lambda}v^* = v^*A^* = v^*A^\top$  and that  $\lambda + \bar{\lambda} = 2\operatorname{Re}(\lambda)$ . Since  $Q$  and  $P$  are positive definite, we conclude that  $\operatorname{Re}(\lambda) < 0$ , and that  $A$  must be Hurwitz.  $\square$

Above, we showed that a matrix  $A \in \mathbb{R}^{n \times n}$  is Hurwitz if and only if the CTLE defined by  $(A, Q)$  has a solution for *all*  $Q \succ 0$ . Is it sufficient to exhibit a solution for a *single*  $Q \succ 0$ ? The following result answers this in the affirmative.

**Corollary 3.5 (Single  $Q$  CTLE Test for Exponential Stability)** *Consider a matrix  $A \in \mathbb{R}^{n \times n}$  and a matrix  $Q \in \mathbb{S}^n$ ,  $Q \succ 0$ .  $A$  is Hurwitz if and only if there exists a solution  $P \in \mathbb{S}^n$ ,  $P \succ 0$  to the continuous-time Lyapunov equation,  $A^\top P + PA = -Q$ .*

**Proof** Suppose there exists a solution  $P \succ 0$  to the Lyapunov equation,

$$A^\top P + PA = -Q. \quad (3.170)$$

Then, using the same eigenvalue argument as in the proof of Theorem 3.9 (pick an eigenvalue, eigenvector pair  $(\lambda, v)$  of  $A$  and left and right multiply the CTLE by  $v^*, v$ ), it follows that  $A$  is Hurwitz.

Now, we show the reverse direction. Suppose  $A$  is Hurwitz. Then, by Theorem 3.9, there exists a solution  $P \succ 0$  to the CTLE  $A^\top P + PA = -Q$  for any  $Q \succ 0$ . Taking  $Q$  to be the matrix in the statement of the corollary, the result follows.  $\square$

Thus, in order to prove that  $x_e = 0$  is globally exponentially stable for  $\dot{x} = Ax$ , it's sufficient to exhibit a solution  $P \succ 0$  to the CTLE  $A^\top P + PA = -Q$  for *any* fixed  $Q \succ 0$ .

**Exercise 3.19** Confirm that  $A$  is Hurwitz if and only if there exists a solution  $P \succ 0$  to the inequality  $A^\top P + PA \prec 0$ .

**Exercise 3.20** Using Theorem 3.9, Corollary 3.5, and their proofs, confirm that the continuous-time conjectures posed above are true.

### 3.2.4.2 The Discrete-Time Lyapunov Equation

Now, let's prove the analogous results in the discrete-time case. In the remainder of this section, we'll prove that positive definite solutions  $P \succ 0$  to the discrete-time Lyapunov equation,  $A^\top P A - P = -Q$ ,  $Q \succ 0$ , completely characterize the exponential stability of discrete-time systems.

**Theorem 3.10 (DTLE Test for Exponential Stability)**  *$A \in \mathbb{R}^{n \times n}$  is Schur if and only if, for any  $Q \in \mathbb{S}^n$ ,  $Q \succ 0$ , there exists a unique solution  $P \in \mathbb{S}^n$ ,  $P \succ 0$  to the DTLE,*

$$A^\top P A - P = -Q. \quad (3.171)$$

Further, the unique solution to the DTLE is computed,

$$P = \sum_{k=0}^{\infty} (A^\top)^k Q A^k. \quad (3.172)$$

*Remark 3.24* Recall that the equilibrium  $x_e = 0$  of a discrete-time system  $x[k+1] = Ax[k]$  is globally exponentially stable if and only if  $A$  is *Schur*—that is, every eigenvalue  $\lambda$  of  $A$  satisfies  $|\lambda| < 1$ . This result therefore characterizes global exponential stability via the discrete-time Lyapunov equation.

**Proof** We follow a strategy similar to the continuous-time case. First, we show that the forward direction holds for any  $Q \in \mathbb{R}^{n \times n}$ . Suppose  $A$  is Schur, and fix a  $Q \in \mathbb{R}^{n \times n}$ . We will show that the proposed formula for  $P$  does indeed solve the DTLE. First, we note that the infinite series for  $P$  converges since  $A$  is Schur. Now, we show that it satisfies the given equation. We have,

$$A^\top P A - P = A^\top \left( \sum_{k=0}^{\infty} (A^\top)^k Q A^k \right) A - \sum_{k=0}^{\infty} (A^\top)^k Q A^k \quad (3.173)$$

$$= \left( \sum_{k=0}^{\infty} (A^{k+1})^\top Q A^{k+1} \right) - \sum_{k=0}^{\infty} (A^\top)^k Q A^k \quad (3.174)$$

$$= -(A^\top)^0 Q A^0 = -Q. \quad (3.175)$$

We conclude that  $P$  satisfies the DTLE. Now, we show that it is the unique solution. We use the same argument as the continuous case. Define an operator  $\Pi : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}^{n \times n}$  by

$$\Pi(X) = A^\top X A - X. \quad (3.176)$$

Above, we showed that this operator is surjective onto  $\mathbb{R}^{n \times n}$ . Since it is also linear in  $X$ , we conclude its kernel equals  $\{0\}$  and that it is invertible. Any solution to the DTLE must therefore be unique.

Now, we confirm that, when  $Q \succ 0$ ,  $P \succ 0$ . Fix a matrix  $Q \succ 0$ . Then, the unique solution  $P$  to the DTLE satisfies  $x^\top P x = \sum_{k=0}^{\infty} x^\top (A^k)^\top Q A^k x$ . Since each term in this sum is nonnegative for  $x \neq 0$ , and the  $k = 0$  term is positive for  $x \neq 0$ , we conclude that  $x^\top P x > 0$  for all  $x \neq 0$ . Since  $P$  is also symmetric, we conclude that  $P \succ 0$ .

Now, we proceed in the opposite direction. Suppose for all  $Q \succ 0$ , there exists a solution  $P \succ 0$  to the DTLE. Fix a  $Q, P$  pair. Let  $\lambda$  be an eigenvalue of  $A$ , with corresponding eigenvector  $v$ . Then,

$$v^* A^\top P A v - v^* P v = -v^* Q v \quad (3.177)$$

$$\bar{\lambda} \lambda v^* P v - v^* P v = -v^* Q v \quad (3.178)$$

$$(|\lambda|^2 - 1) v^* P v = -v^* Q v \quad (3.179)$$

$$|\lambda|^2 \leq 1 - \frac{v^* Q v}{v^* P v}. \quad (3.180)$$

Since  $Q, P \succ 0$ , we conclude that  $|\lambda|^2 < 1$ , and that  $A$  must be Schur.  $\square$

We finish up by stating a discrete-time analogue of Corollary 3.5.

**Corollary 3.6 (Single  $Q$  DTLE Test for Exponential Stability)** *Consider a matrix  $A \in \mathbb{R}^{n \times n}$  and a matrix  $Q \in \mathbb{S}^n, Q \succ 0$ .  $A$  is Schur if and only if there exists a solution  $P \in \mathbb{S}^n, P \succ 0$  to the discrete-time Lyapunov equation,  $A^\top P A - P = -Q$ .*

**Exercise 3.21** Prove Corollary 3.6.

**Exercise 3.22** Confirm that  $A$  is Schur if and only if there exists a solution  $P \succ 0$  to the inequality  $A^\top P A - P \prec 0$ .

**Exercise 3.23** Using Theorem 3.10, Corollary 3.6, and their proofs, confirm that the discrete-time conjectures posed above are true.

### 3.2.5 Further Reading

Exercise 3.15, concerning evaluation of eigenvalues of a Jordan form by inspection, is from [40]. The treatment of phase portraits of planar systems closely follows [31]. The treatment of Lyapunov theory in this section relied on algebraic techniques largely to avoid introducing the comparison lemma in the continuous-time case—a comparison lemma-based treatment can be found in [20]. The development of the Lyapunov equations for the Lyapunov and exponentially stable cases is based on the those of [36] and [14].

### 3.2.6 Problems

**Problem 3.6 (Distinct Eigenvalues [9])** Consider the linear system  $\dot{x} = Ax$ , where  $A \in \mathbb{R}^{n \times n}$  has  $n$  linearly independent eigenvectors  $v_1, \dots, v_n$  with corresponding (not necessarily distinct) eigenvalues  $\lambda_1, \dots, \lambda_n$ . Find  $x(t)$  in terms of the eigenvalues and eigenvectors of  $A$ , and give a geometric interpretation of the result when  $x_0 = v_i, i = 1, \dots, n$ . Describe how to compute the stable, unstable, and center subspaces from this information.

**Problem 3.7 (A Geometric Proof of the Invariant Subspace Theorem)** Above, we used the real Jordan form to prove the invariant subspace theorem. In this problem, we'll give a *geometric* proof of this theorem. Consider a matrix  $A \in \mathbb{R}^{n \times n}$ , and a continuous-time, LTI system  $\dot{x} = Ax$ .

1. Show that each generalized eigenspace  $K_\lambda(A)$  is invariant under  $A$ . That is,  $v \in K_\lambda(A)$  implies  $Av \in K_\lambda(A)$ .
2. Using your answer to part (1), show that each generalized eigenspace  $K_\lambda(A)$  is invariant under  $\exp(At)$  for all  $t$ . That is,  $v \in K_\lambda(A)$  implies  $\exp(At)v \in K_\lambda(A)$  for all  $t \in \mathbb{R}$ . *Hint: are finite-dimensional normed vector spaces complete?*
3. Using your answer to part (2), prove the invariant subspace theorem.

**Problem 3.8 (An Analytical Approach to the Lyapunov Equation)** In this problem, we take a strictly analytical approach to deriving the Lyapunov equation results we posed above. Here, we'll focus primarily on the continuous-time case, as the theory is somewhat more delicate.

1. First, we'll prove a *linear comparison lemma*. Let  $x : \mathbb{R} \rightarrow \mathbb{R}$  be a  $C^1$  function satisfying,

$$\dot{x}(t) \leq \lambda x(t), \quad (3.181)$$

for some constants  $\lambda \in \mathbb{R}$ . Show that if  $y : \mathbb{R} \rightarrow \mathbb{R}$  satisfies  $\dot{y}(t) = \lambda y(t)$  for all  $t \in \mathbb{R}$ , and  $x(t_0) \leq y(t_0)$  for some  $t_0 \in \mathbb{R}$ , then  $x(t) \leq y(t)$  for all  $t \geq t_0$ .

2. Suppose there exists a function  $V : \mathbb{R}^n \rightarrow \mathbb{R}$  and constants  $c_i > 0$ ,  $i = 1, 2, 3$ , for which

$$c_1 \|x\|^2 \leq V(x) \leq c_2 \|x\|^2 \quad (3.182)$$

$$\frac{\partial V}{\partial x} Ax \leq -c_3 \|x\|^2, \quad (3.183)$$

for all  $x \in \mathbb{R}^n$ . Using part (1), prove that the equilibrium  $x_e = 0$  of  $\dot{x} = Ax$  is globally exponentially stable.

3. Fix a positive definite matrix  $Q \succ 0$ . Using parts (1) and (2), prove that the equilibrium  $x_e = 0$  of  $\dot{x} = Ax$  is globally exponentially stable if there exists a positive definite solution  $P \succ 0$  of the CTLE,  $A^\top P + PA = -Q$ .

**Problem 3.9 (The Sylvester Equation ★)** In the previous section, we examined the continuous-time *Lyapunov equation*. This equation is an instance of a more general type of matrix equation called a *Sylvester equation*. A matrix equation of the form,

$$MX + XN = Q, \quad (3.184)$$

in an unknown matrix  $X$ , is said to be a Sylvester equation in  $X$ . In this problem, we'll analyze the set of Sylvester equations using an algebraic approach different to that which we've considered so far.

1. For matrices  $M \in \mathbb{R}^{n \times n}$  and  $N \in \mathbb{R}^{n \times n}$ , the Kronecker product of  $M$  and  $N$ , denoted  $M \otimes N$ , is the matrix

$$M \otimes N := \begin{bmatrix} m_{11}N & \dots & m_{1n}N \\ \vdots & \ddots & \vdots \\ m_{n1}N & \dots & m_{nn}N \end{bmatrix}. \quad (3.185)$$

Let  $X = [x_1, \dots, x_n] \in \mathbb{R}^{n \times n}$  and  $Q = [q_1, \dots, q_n] \in \mathbb{R}^{n \times n}$ , where each  $x_i$  and  $q_i$  denotes a column of  $X$  and  $Q$ , respectively. Define,

$$\text{vec}(X) = [x_1^\top \dots x_n^\top]^\top \in \mathbb{R}^{n^2} \quad (3.186)$$

$$\text{vec}(Q) = [q_1^\top \dots q_n^\top]^\top \in \mathbb{R}^{n^2}. \quad (3.187)$$

Using the Kronecker product, prove there exists a matrix  $A \in \mathbb{R}^{n^2 \times n^2}$  for which

$$A \text{vec}(X) = \text{vec}(Q) \iff MX + XN = Q. \quad (3.188)$$

2. Suppose the eigenvalues of  $M$  are  $\lambda_1, \dots, \lambda_n$  and the eigenvalues of  $N$  are  $\mu_1, \dots, \mu_n$ . Show that the eigenvalues of  $M \otimes N$  are the  $n^2$  numbers  $\lambda_i \mu_j$ ,  $(i, j) \in \{1, \dots, n\}^2$ .
3. Using your answers to parts (1) and (2), show there exists a unique solution  $X$  to the Sylvester equation  $MX + XN = Q$  if and only if  $\lambda_i(M) + \lambda_j(N) \neq 0$  for all  $(i, j) \in \{1, \dots, n\}^2$ . Specialize this result to conclude the CTLE test for a Hurwitz matrix.

**Problem 3.10 (Lyapunov Certificates of Robust Stability)** When analyzing the stability of linear, time-invariant systems, we assumed that we had *perfect* knowledge of the matrix  $A$  determining the system's dynamics. In this problem, we'll introduce *uncertainty* into the system, and show that we can use the Lyapunov equation to get a certificate of *robust stability*. Consider an uncertainty set  $\Delta$ , defined as the convex hull of a finite collection  $\{\Delta_1, \dots, \Delta_m\}$ , where  $\Delta_i \in \mathbb{R}^{n \times n}$ ,

$$\Delta = \text{conv}\{\Delta_1, \dots, \Delta_m\} = \left\{ \sum_{i=1}^m \lambda_i \Delta_i : \sum_{i=1}^m \lambda_i = 1, 0 \leq \lambda_i \leq 1, \lambda_i \in \mathbb{R} \right\} \subseteq \mathbb{R}^{n \times n}. \quad (3.189)$$

Consider an unforced, uncertain system  $\dot{x}(t) = (A + \Delta)x(t)$ , where  $\Delta \in \Delta$  is some unknown parameter. If  $x_e = 0$  is exponentially stable for  $\dot{x}(t) = (A + \Delta)x(t)$  for all  $\Delta \in \Delta$ , we say that  $x_e = 0$  is *robustly exponentially stable*. In this problem, we'll find a certificate for robust exponential stability using *linear matrix inequalities* (LMIs). Let  $\mathbb{S}^n$  represent the set of  $n \times n$  real, symmetric matrices and  $V$  represent a finite-dimensional vector space. A linear matrix inequality in  $X \in V$  is a matrix inequality of the form  $F(X) \preceq Q$ , where  $F : V \rightarrow \mathbb{S}^n$  is a linear map,  $Q \in \mathbb{S}^n$ , and  $F(X) \preceq Q$  denotes the positive semidefinite constraint  $0 \preceq Q - F(X)$ . The linear matrix inequality  $F(X) \preceq Q$  has a solution if there exists an  $X \in V$  for which  $F(X) \preceq Q$ .

1. Let's begin by synthesizing an LMI to identify a Lyapunov function for the unforced system  $\dot{x}(t) = Ax(t)$  without uncertainty. Show that  $A$  is Hurwitz if and only if the pair of linear matrix inequalities,

$$I \preceq P, \quad A^\top P + PA \preceq -I, \quad (3.190)$$

in a symmetric matrix  $P \in \mathbb{S}^n$ , has a solution.

2. Now, let's introduce the uncertainty set  $\Delta = \text{conv}\{\Delta_1, \dots, \Delta_m\}$ . Prove there exists a finite collection of linear matrix inequalities for which the existence of a solution to the inequalities implies the origin of  $\dot{x} = (A + \Delta)x$ ,  $\Delta \in \Delta$  is robustly exponentially stable.
3. Now, let's introduce control into the system. Consider the uncertain state equation,

$$\dot{x} = (A + \Delta)x + Bu, \quad (3.191)$$

where  $u \in \mathbb{R}^m$  is a control input and  $\Delta \in \Delta = \text{conv}\{\Delta_1, \dots, \Delta_m\}$ . Devise a linear matrix inequality method for computing a matrix  $K \in \mathbb{R}^{m \times n}$  for which the control law  $u = Kx$  renders the equilibrium  $x_e = 0$  of  $\dot{x} = (A + \Delta)x + Bu$  robustly exponentially stable. Produce an example of  $(A, B, \Delta)$  for which your method will be successful, and an example for which it will be unsuccessful. *Hints: Your LMIs do not have to directly give you  $K$ . How are the eigenvalues of  $A$  and  $A^\top$  related?*

**Problem 3.11 (Stable Subspaces [27])** Consider the linear system  $\dot{x} = Ax$ , where the matrix  $A$  is given by

$$A = \begin{bmatrix} \lambda_1 & 1 \\ 0 & \lambda_2 \end{bmatrix}, \quad \lambda_1, \lambda_2 \in \mathbb{R}. \quad (3.192)$$

1. Find the stable, unstable, and center subspaces  $E^s$ ,  $E^u$ , and  $E^c$  for  $\lambda_1 > 0$  and  $\lambda_2 < 0$ .
2. The phase portrait of the system is a plot of  $x_1$  versus  $x_2$  trajectories. Qualitatively sketch the phase portrait of the system:

- a. For  $\lambda_1, \lambda_2 > 0$ .
  - b. For  $\lambda_1, \lambda_2 < 0$ .
  - c. For  $\lambda_1 > 0$  and  $\lambda_2 < 0$ .
3. Compute the matrix exponential,  $\exp(At)$ , for all  $\lambda_1, \lambda_2 \in \mathbb{R}$ .
  4. From part (a), verify that  $\mathbb{R}^2 = E^s \oplus E^u \oplus E^c$ , and that these subspaces are invariant under  $\exp(At)$ .

**Problem 3.12 (Degenerate Critical Points & Phase Portraits [31])** Above, we examined the phase portraits of systems where the *origin* is the only equilibrium point. If the origin is *not* the only equilibrium, it is referred to as the “degenerate critical point” of  $\dot{x} = Ax$ . In each of the following cases, determine the solution and the corresponding phase portraits:

$$A_1 = \begin{bmatrix} \lambda & 0 \\ 0 & 0 \end{bmatrix}, A_2 = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, A_3 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}. \quad (3.193)$$

Note that the origin is *not* an isolated equilibrium point in these cases. The four different phase portraits determined by  $A_1$  (for  $\lambda > 0$  or  $\lambda < 0$ ) and  $A_2, A_3$ , together with the examples discussed in the section, compose the eight different types of qualitative behavior that are possible for a continuous-time, LTI system.

**Problem 3.13 (The Discrete-Time Invariant Subspace Theorem)** State and prove a discrete-time analogue of the invariant subspace theorem. Make sure to highlight any additional assumptions or pieces that change from the continuous-time case.

**Problem 3.14 (Some Special Lyapunov Equations [9])** In this problem, we’ll analyze some special cases of the continuous-time Lyapunov equation. Consider the continuous-time, SISO LTI system,  $\dot{x} = Ax + Bu$ ,  $y = Cx$ , for which  $x \in \mathbb{R}^n$ ,  $u \in \mathbb{R}$ , and  $y \in \mathbb{R}$ .

1. Suppose there exists a positive definite matrix  $P \in \mathbb{S}^n$ ,  $P \succ 0$ , and a constant  $\alpha \in \mathbb{R}$ , for which

$$A^\top P + PA \prec \alpha P. \quad (3.194)$$

Which region in the complex plane do the eigenvalues of  $A$  lie in? Write your answer in terms of  $\alpha$ .

2. Let’s design a basic feedback controller using the result of part (1). Suppose Equation (3.194) holds with  $\alpha = 0$ , and in addition, that  $PB = C^\top$ . Show that, for a feedback control law  $u = -ky$ ,  $k \geq 0$ , the equilibrium  $x_e = 0$  of the closed-loop system  $\dot{x} = Ax - Bky$  is globally exponentially stable.
3. Now, suppose  $A \in \mathbb{R}^{n \times n}$  and  $P \in \mathbb{S}^n$ ,  $P \succ 0$  satisfy the equality  $A^\top P + PA = 0$ . Which region of the complex plane do the eigenvalues of  $A$  lie in?
4. Assuming again that  $PB = C^\top$ , does the equality  $A^\top P + PA = 0$  still guarantee exponential stability for a control law  $u = -ky$ ,  $k \geq 0$ ? If not, what additional conditions would you need?

**Problem 3.15 (Lyapunov Ellipsoid Identification [6])** Consider an unforced, LTI system  $\dot{x}(t) = Ax(t)$ ,  $x \in \mathbb{R}^n$ . A forward invariant set for this system is a set  $\mathcal{S} \subseteq \mathbb{R}^n$  for which  $x_0 \in \mathcal{S}$  implies  $\varphi(t, t_0, x_0) \in \mathcal{S}$  for all  $t \geq t_0$ , where  $\varphi$  is the (unforced) state transition map. In this problem, we’ll develop a technique for certifying the invariance of certain sets.



1. An ellipsoid in  $\mathbb{R}^n$ , centered at zero, can be defined in two ways:

$$\mathcal{E}_1 = \{x \in \mathbb{R}^n : x^\top S x \leq 1\}, \quad \text{where } S \succ 0, \quad (3.195)$$

$$\mathcal{E}_2 = \{x \in \mathbb{R}^n : x = My, \|y\|_2 \leq 1\}, \quad \text{where } \det M \neq 0. \quad (3.196)$$

Given  $S$  for  $\mathcal{E}_1$ , explain how to find  $M$  for  $\mathcal{E}_2$  such that  $\mathcal{E}_1 = \mathcal{E}_2$ . Given  $M$  for  $\mathcal{E}_2$ , explain how to find  $S$  for  $\mathcal{E}_1$  such that  $\mathcal{E}_1 = \mathcal{E}_2$ . Support your claims with proofs.

2. Using a collection of linear matrix inequalities, explain how to find an ellipsoid  $\mathcal{E} \subseteq \mathbb{R}^n$ , centered at 0, that is forward invariant for  $\dot{x} = Ax$  and satisfies  $z^{(i)} \in \mathcal{E}$ ,  $i = 1, \dots, m$  and  $w^{(j)} \notin \mathcal{E}$ ,  $j = 1, \dots, p$ , for  $z^{(i)}, w^{(j)}$  some fixed points in  $\mathbb{R}^n$ . *Note: you may leave your solution in terms of strict inequalities (e.g.  $\prec, <$ ) for the purposes of this problem. For computer implementation, strict inequalities are not realizable. Think about how you can re-pose your problem with non-strict inequalities!*
3. Produce a simple example for which your method from part (2) has a solution, and an example for which it does not. You may present your answer in the form of a simple sketch or in the form of problem data.

**Problem 3.16 (Constant Norm & Constant Speed Systems [6])** The system  $\dot{x} = Ax$  is called *constant norm* if, for every trajectory  $x$ ,  $\|x(t)\|$  is constant. The system is called *constant speed* if for every trajectory  $x$ ,  $\|\dot{x}(t)\|$  is constant.

1. Find the (general) conditions on  $A$  under which the system is constant norm.
2. Find the (general) conditions on  $A$  under which the system is constant speed.
3. Is every constant norm system a constant speed system? Provide a proof or counterexample.
4. Is every constant speed system a constant norm system? Provide a proof or counterexample.

**Problem 3.17 (Peaking Factor [6] ★)** Consider the linear, time-invariant system  $\dot{x} = Ax$ , where  $x \in \mathbb{R}^n$ . We define the *peaking factor*  $p$  of the system as,

$$p := \sup_{t \geq 0, x(0) \neq 0} \frac{\|x(t)\|}{\|x(0)\|}, \quad (3.197)$$

where the supremum is taken over all trajectories of the system with  $x(0) \neq 0$ .

1. Prove it is always true that  $p \geq 1$ . Then, show that  $p = \infty$  implies  $x_e = 0$  is unstable.
2. Suppose the system has a Lyapunov function  $V(x) = x^\top Px$ , for  $P \succ 0$  such that  $\dot{V}(x) \leq 0$  for all  $x \in \mathbb{R}^n$ . Show that the peaking factor satisfies  $p \leq \sqrt{\kappa}$ , where  $\kappa = \lambda_{\max}(P)/\lambda_{\min}(P)$  is the condition number of  $P$ .

### 3.3 Norms on Signals & Systems

Thus far, we've exclusively focused on the *state space* perspective on stability, examining the Lyapunov, asymptotic, and exponential stability of unforced linear systems. In this section, instead of taking this *internal* perspective, we take an *external* perspective and study input/output stability. By defining appropriate notions of norms on signals and systems, we'll come up with a theory of stability that characterizes how the size of an input signal affects the size of an output signal. We'll find that transfer functions give us a natural language to discuss these concepts, and will develop some more advanced transfer function theory along the way.

Let's sketch out a basic framework for input/output stability theory. Recall that, in Chapter 2, we showed that the zero-state response of a linear, time-invariant system with impulse response map  $G$  is computed via a convolution,

$$y(t) = (G * u)(t) \quad (3.198)$$

$$y[k] = (G * u)[k]. \quad (3.199)$$

The basic idea of I/O stability is the following: by treating convolution with  $G$  as an *operator* and finding a norm on that operator, we can estimate how a system amplifies an input signal and therefore make claims about its input/output stability.

Let's develop a few more details of this idea. First, we know that convolution with  $G$  is a *linear* operation—for all admissible signals  $u$  (with a well-defined convolution), one has  $(G * (\alpha u_1 + \beta u_2)) = \alpha G * u_1 + \beta G * u_2$ . Thus,  $G*$  can be thought of as a linear operator from a subspace of  $\mathcal{U}$ , having a well-defined convolution, to  $\mathcal{U}$ . Once we take this perspective, we can compute the induced operator norm of  $G*$  just like we can for any other linear operator. By choosing norms  $\|\cdot\|_{\mathcal{U}}$  and  $\|\cdot\|_{\mathcal{Y}}$  on the input and output spaces, we can identify the induced operator norm of  $G*$  by finding the “smallest” constant  $C \geq 0$  for which,

$$\|y\|_{\mathcal{Y}} = \|G * u\|_{\mathcal{Y}} \leq C \|u\|_{\mathcal{U}}, \quad \forall u(\cdot) \in \mathcal{U}. \quad (3.200)$$

where  $\|\cdot\|$  is some norm on the vector space of signals. How might we actually compute the induced norm for a general linear, time-invariant system? Perhaps surprisingly, what we'll find is that, the norm of  $G*$  in the time domain can be computed from a norm on its corresponding transfer function  $\hat{G}$  in the frequency domain! This finding is indicative of a larger theme that we'll encounter across the remainder of the course: that time domain specifications of performance manifest themselves in frequency domain conditions on a transfer function. As a consequence of this connection, it behooves us to learn a little bit more transfer function theory before taking on the problem of input/output stability.

#### 3.3.1 Poles & Zeros of Transfer Functions

As mentioned above, in order to understand how a linear, time-invariant system amplifies its input signals, it's useful to develop a more nuanced understanding of transfer functions. As the few of the results we'll examine here require a (somewhat advanced) knowledge of complex analysis (something we do not assume the reader possesses), we'll forego several of

the proofs in this subsection. For each such result, we'll provide references in which proofs can be found. With this said, let's get started!

First, we recall a little bit of the transfer function theory that we developed in Chapter 2. You might remember that, when computing the transforms of various common signals, we frequently encountered expressions such as,

$$\frac{1}{s+1}, \frac{z+2}{(z+1)(z+3)}, \dots, \text{etc.} \quad (3.201)$$

Such functions are said to be *rational functions* of  $s$  and  $z$ , as they are the ratios of two polynomials in  $s$  and  $z$ . What else did we find? We also showed that the transfer functions of continuous and discrete-time LTI system representations are computed,

$$\hat{G}(s) = C(sI - A)^{-1}B + D \quad (3.202)$$

$$\hat{G}(z) = C(zI - A)^{-1}B + D. \quad (3.203)$$

Are these also rational functions of  $s$  and  $z$ ? We recall the fundamental fact from linear algebra that, for an invertible matrix  $A \in \mathbb{R}^{n \times n}$ , the entries of  $A^{-1}$  consist of rational combinations of the entries of  $A$ . Thus, the entries of the transfer functions  $\hat{G}(s)$  and  $\hat{G}(z)$  must *also* be rational functions of  $s$  and  $z$ . It therefore seems that, for a wide variety of “interesting” signals and systems, rational functions are the “right” class of function for transforms. In order to develop a more sophisticated understanding of transforms, we'll therefore undertake a brief study of rational functions. We begin by providing a formal definition of a class of rational functions.

**Definition 3.14 (Real-Rational Function)** Let  $\Omega \subseteq \mathbb{C}$  be a domain with nonempty interior. A scalar function  $f : \Omega \rightarrow \mathbb{C}$  is said to be a real-rational function if,

$$f(s) = \frac{n(s)}{d(s)}, \quad \forall s \in \Omega, \quad (3.204)$$

where  $n : \mathbb{C} \rightarrow \mathbb{C}$  and  $d : \mathbb{C} \rightarrow \mathbb{C}$  are polynomials with real coefficients,

$$n(s) = b_m s^m + \dots + b_0, \quad b_i \in \mathbb{R}, \quad (3.205)$$

$$d(s) = a_n s^n + \dots + a_0, \quad a_i \in \mathbb{R}, \quad (3.206)$$

and  $d$  is not the zero polynomial. A matrix-valued function  $F : \Omega \rightarrow \mathbb{C}^{p \times k}$  is said to be a real-rational matrix function if each element  $[F]_{ij} : \Omega \rightarrow \mathbb{C}$  is a real-rational function.

*Remark 3.25* By convention, we take the domain  $\Omega \subseteq \mathbb{C}$  of a real-rational function  $f(s) = n(s)/d(s)$  to be  $\mathbb{C}$  minus the roots of the denominator polynomial  $d(s)$ . For the case of a matrix-valued, real-rational function, we take  $\Omega$  to be  $\mathbb{C}$  minus the roots of each entry's denominator polynomial.

*Remark 3.26* In abstract algebra, it's common to think of real-rational functions not as *functions* which we can plug numbers into, but rather elements of an abstract algebraic structure called a *ring*. In this paradigm, one doesn't consider  $s$  to be a variable, but rather a special polynomial called the *indeterminate*. In this setting, since  $s$  is not thought of as something that can be “substituted in for,” the choice of domain  $\Omega$  is typically not of concern. Since this is a degree of abstraction beyond what's needed in this course, we'll stick with treating real-rational functions simply as standard functions on a domain  $\Omega$ .

Let's examine the basic structure of these real-rational functions. From the fundamental theorem of algebra, we know that we can factor the numerator and denominator polynomials of a real-rational function as,

$$f(s) = \frac{b_ms^m + \dots + b_0}{a_ns^n + \dots + a_0} = \frac{(s - z_1) \cdot \dots \cdot (s - z_m)}{(s - r_1) \cdot \dots \cdot (s - r_n)}, \quad (3.207)$$

where  $z_1, \dots, z_m \in \mathbb{C}$  are the  $m$  (potentially repeated) roots of the numerator polynomial and  $r_1, \dots, r_n \in \mathbb{C}$  are the  $n$  (potentially repeated) roots of the denominator polynomial. When studying transfer functions in the context of stability, we'll be particularly interested in the case where the numerator and denominator polynomials have no common roots. In this case, the two polynomials are said to be *coprime*.

**Definition 3.15 (Coprime Polynomials)** Two polynomials  $p, q : \mathbb{C} \rightarrow \mathbb{C}$  with real coefficients are said to be coprime if they have no common roots.

As a consequence of the factorization of a real-rational function  $f(s)$ , one can always simplify a real-rational function into a form where its numerator and denominator polynomials are coprime by cancelling the common roots of the numerator and denominator polynomials. We define this “simplest form,” in which all common roots have been cancelled, as the *coprime form* of a rational function.

**Definition 3.16 (Coprime Form)** Consider a real-rational function  $f : \Omega \rightarrow \mathbb{C}$ . The *coprime form* of  $f$ , denoted  $f_c$ , is the real-rational function

$$f_c(s) = \frac{n_c(s)}{d_c(s)}, \quad (3.208)$$

where  $n_c$  and  $d_c$  are coprime polynomials and  $f_c$  satisfies  $f_c(s) = f(s) \forall s \in \Omega$ . If  $F : \Omega \rightarrow \mathbb{C}^{p \times k}$  is a real-rational function, then its coprime form  $F_c$  is the real-rational function containing the coprime of  $[F_c]_{ij}$  in each entry.

*Remark 3.27* The *coprime form* is referred to by some authors as the *canonical fractional representative*. Here, we'll stick with *coprime form*, as it's somewhat easier to digest.

Thus, the coprime form of a real-rational function  $f$  is nothing more than the “simplest form” of  $f$ , where all like factors in the numerator and denominator have been cancelled. For instance, the coprime form of  $f(s) = (s + 1)/(s + 1)^2$  is,

$$f_c(s) = \frac{1}{s + 1}. \quad (3.209)$$

In order to ensure that the coprime form is well-defined, we must check that it exists *and* is unique for any given real-rational function. If these conditions can be verified, it will make sense to talk about *the* coprime form of a real-rational function.

**Theorem 3.11 (Existence & Uniqueness of Coprime Form)** *For any scalar real-rational function  $f : \Omega \rightarrow \mathbb{C}$ , there exist unique coprime, monic<sup>3</sup> polynomials  $n_c : \mathbb{C} \rightarrow \mathbb{C}$  and  $d_c : \mathbb{C} \rightarrow \mathbb{C}$ , and a constant  $c \in \mathbb{R}$ , for which*

<sup>3</sup> A polynomial is said to be *monic* if the coefficient of its highest degree term is equal to 1. For example, the polynomial  $s^n + a_{n-1}s^{n-1} + \dots + a_0$  is a monic polynomial.

$$f(s) = c \frac{n_c(s)}{d_c(s)}, \forall s \in \Omega. \quad (3.210)$$

Thus, the coprime form of a real-rational function is well-defined.

**Exercise 3.24** Verify the claim of Theorem 3.11 using the fundamental theorem of algebra. Recall that, for a polynomial with real coefficients, complex roots come in conjugate pairs.

Now, we discuss what it means for two real-rational functions to be equal. Consider the real-rational functions,

$$f_1(s) = \frac{(s+1)}{(s+2)(s+1)}, \quad f_2(s) = \frac{1}{(s+2)}. \quad (3.211)$$

Clearly, if one is willing to cancel the factor of  $s+1$  in the numerator and denominator of  $f_1(s)$ , one will get the polynomial  $f_2(s)$ . Does this mean that  $f_1$  and  $f_2$  are equal? In the algebraic study of rational functions, the answer is *yes*. This is the convention we will adopt in our study of transfer functions.

**Definition 3.17 (Equivalent Real-Rational Functions)** Two real-rational functions  $f_1(s)$  and  $f_2(s)$  are said to be equivalent, written  $f_1(s) = f_2(s)$ , if they have the same coprime form.

*Remark 3.28* The notation  $f_1(s) = f_2(s)$  should not be interpreted as an equality in the standard, functional sense. Here, when we write equality, we mean that  $f_1(s)$  and  $f_2(s)$  *represent* the same rational function—not that they are truly equal to each other for all values of  $s$ . In the example above, for instance,  $f_1$  is not defined for  $s = -1$ , but  $f_2$  is, which means that  $f_1$  and  $f_2$  cannot be equal as *functions* in the standard sense. However, under simplification to coprime form, they *represent* the same rational functions, so we write  $f_1(s) = f_2(s)$ . In order to formalize this notion of “equivalence,” and write this equality in a mathematically rigorous sense, one can define special sets called *equivalence classes* on the set of real-rational functions. We’ll avoid explicitly using this formality for ease of exposition—we direct the interested reader to Problem 3.18 for more details.

In light of Definition 3.17, we consider the real-rational functions,

$$\frac{(s+1)}{(s+1)^2}, \quad \frac{(s+1)(s+2)}{(s+1)^2(s+2)}, \quad \frac{1}{s+1}, \quad (3.212)$$

to be equivalent, as each function has the coprime form  $1/(s+1)$ .

Let’s again focus in on the case of a transfer function  $\hat{G}(s) = C(sI - A)^{-1}B + D$ . What additional structure exists in real-rational functions of this form? As another consequence, of the definition of the matrix inverse, one may show that the degree of the denominator polynomial(s) associated to  $\hat{G}(s)$  is always greater than or equal to that of the numerator polynomial. Further, when  $D = 0$ , this inequality becomes *strict*. This leads us to the following definition.

**Definition 3.18 (Proper/Strictly Proper Real-Rational Function)** Consider a scalar, real-rational function  $f(s) = n(s)/d(s)$  in coprime form.

1. Proper: the function is *proper* if  $\deg(d) \geq \deg(n)$ .
2. Strictly Proper: the function is *strictly proper* if  $\deg(d) > \deg(n)$ .

A real-rational matrix function  $F$  is said to be (strictly) proper if each of its entries is (strictly) proper. A real-rational function that is not proper is said to be *improper*.

*Remark 3.29* Above, we stated a definition of *proper* and *strictly proper* for real-rational functions in coprime form. Under our convention for equivalence, any real-rational function is equivalent to its coprime form. Thus, the properness of any real-rational function is simply determined by the properness of its coprime form. We'll continue stating definitions for real-rational functions in coprime form as we proceed through this section—keep in mind that all of these definitions are fully general under our convention for equivalence! Simply reduce your function to coprime form first, and then apply the definition.

Now that we've laid out the basic properties of real-rational functions, we can examine some more system-theoretic concepts. Given a transfer function  $\hat{G}$  for a causal, LTI system, it seems reasonable to ask: when is the function zero, when does it “blow up,” and what is the significance of these zero and “blow up” points? First, we answer these questions for the SISO case, in which the transfer function is a scalar, real-rational function.

**Definition 3.19 (Poles & Zeros, SISO Case)** Consider a scalar, real-rational transfer function  $\hat{G} : \Omega \rightarrow \mathbb{C}$ ,  $\hat{G}(s) = n(s)/d(s)$ , which is in coprime form.

1. **Zeros:** The zeros of  $\hat{G}$  are the roots of the numerator polynomial,  $n(s)$ .
2. **Poles:** The poles of  $\hat{G}$  are the roots of the denominator polynomial,  $d(s)$ .

*Remark 3.30* Remember—since we consider rational functions to be equal up to coprime simplification, this definition holds even when  $\hat{G}$  is *not* initially in coprime form. One simply needs to cancel out common components of the numerator and denominator, and then calculate the poles and zeros.

*Remark 3.31* We stated this definition using the Laplace transform variable,  $s$ . The definition is *identical* for the  $\mathcal{Z}$ -transform, and  $s$  can be harmlessly swapped out for  $z$ .

*Remark 3.32* Since the coefficients of the polynomials  $n$  and  $d$  are real, poles/zeros are either real *or* appear in complex conjugate pairs. Since we ask that  $\hat{G} = n(s)/d(s)$  be in coprime form, there will be no overlap between the poles and zeros of a SISO transfer function.

Now, we confirm that the poles and zeros of a SISO transfer function do indeed correspond to the “blow up” and zero points of the transfer function.

**Lemma 3.6 (SISO Poles & Zeros)** Consider a scalar, real-rational transfer function  $\hat{G} : \Omega \rightarrow \mathbb{C}$ ,  $\hat{G}(s) = n(s)/d(s)$ . If  $n$  and  $d$  are coprime, then  $s_0$  is a zero of  $\hat{G}$  if and only if

$$\lim_{s \rightarrow s_0} \hat{G}(s) = 0. \quad (3.213)$$

Likewise, if  $n$  and  $d$  are coprime, then  $s_0$  is a pole of  $\hat{G}$  if and only if,

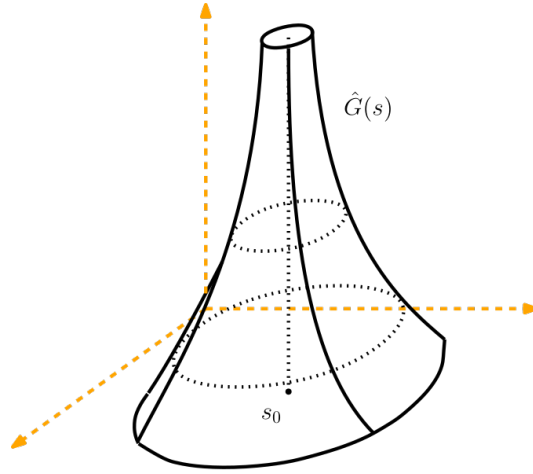
$$\lim_{s \rightarrow s_0} \hat{G}(s) = \infty. \quad (3.214)$$

*Remark 3.33* The alternate characterizations of poles and zeros presented in this lemma are often taken as definitions of poles and zeros for more general, non-rational functions.

*Remark 3.34* A brief word is in order regarding the meaning of  $\infty$  in the complex plane. Since  $\hat{G}$  is a complex function, the meaning of  $\lim_{s \rightarrow s_0} \hat{G}(s) = \infty$  has slightly different meaning than in  $\mathbb{R}$ . By  $\lim_{s \rightarrow s_0} \hat{G}(s) = \infty$ , we mean  $|\hat{G}(s)| \rightarrow \infty$  as  $s \rightarrow s_0$ . For more information about the meaning of  $\infty$  in the complex plane, you can read about the *Riemann sphere* in [38] or any other complex analysis text.

**Proof** Suppose we're given a real-rational transfer function  $\hat{G}(s)$  in coprime form, and a point  $s_0 \in \mathbb{C}$  for which  $\lim_{s \rightarrow s_0} \hat{G}(s) = 0$ . For this limit to go to zero, one must have  $\lim_{s \rightarrow s_0} n(s) = 0$ , which, by continuity of  $n(s)$ , necessitates  $n(s_0) = 0$ . Thus, such an  $s_0$  must be a zero of the transfer function. Now, suppose  $\lim_{s \rightarrow s_0} \hat{G}(s) = \infty$ . Then, it must be that  $\lim_{s \rightarrow s_0} d(s) = 0$ , from which we conclude that  $d(s_0) = 0$ . Thus,  $s_0$  must be a pole of the transfer function.

Now, we prove the converse directions, using that  $n$  and  $d$  are coprime. If  $s_0$  is a zero of  $\hat{G}$ , then  $s_0$  cannot be a root of  $d$ , since  $n$  and  $d$  are coprime. This implies that  $\lim_{s \rightarrow s_0} n(s)/d(s) = n(s_0)/d(s_0) = 0$ . Alternative, suppose  $s_0$  is a pole of  $\hat{G}$ . Then,  $s_0$  is a root of  $d$  and not of  $n$ , which implies  $\lim_{s \rightarrow s_0} d(s)/n(s) = 0$  and  $\lim_{s \rightarrow s_0} \hat{G}(s) = \infty$ .  $\square$



**Fig. 3.8** At a pole, the magnitude of the transfer function will “blow up” to  $\infty$ . The appearance of the graph of the magnitude on  $\mathbb{C}$  lends the pole its name.

Now, we extend the definitions of poles and zeros to the MIMO case. A natural generalization of a scalar being zero to the matrix setting is a matrix *dropping rank*. The notion of a matrix transfer function “dropping rank” at a particular point in  $\mathbb{C}$  yields the most common definition for a MIMO zero. In order to make the idea of a matrix function “dropping rank” precise, we first need to define what the rank of the transfer function normally is.

**Definition 3.20 (Normal Rank)** Let  $\hat{G} : \Omega \rightarrow \mathbb{C}^{p \times m}$  be a real-rational matrix transfer function. The normal rank of  $\hat{G}$  is defined,

$$\text{normalrank}(\hat{G}) = \max_{s \in \Omega} \text{rank}(\hat{G}(s)). \quad (3.215)$$

Thus, the normal rank is the maximum rank of a transfer function across its domain. It's important to stress—since the normal rank is defined by taking the maximum rank over the

entire domain  $\Omega$ , normal rank is a property of the *function*  $\hat{G}$ , not a property of a value  $\hat{G}(s_0)$  as a particular point  $s_0 \in \Omega$ . With this definition in hand, we extend the definitions of poles and zeros to the MIMO case.

**Definition 3.21 (Poles & Zeros, MIMO case)** Consider a real-rational matrix transfer function,  $\hat{G} : \Omega \rightarrow \mathbb{C}^{p \times m}$ , which has been simplified to its coprime form.

1. Transmission zeros: if  $\hat{G}$  has nonzero normal rank, the transmission zeros of  $\hat{G}$  are the points  $s_0 \in \Omega$  at which

$$\text{rank } \hat{G}(s_0) < \text{normalrank}(\hat{G}). \quad (3.216)$$

If  $\hat{G}$  has zero normal rank, then every point of  $\mathbb{C}$  is a zero of  $\hat{G}$ .

2. Poles: the poles of  $\hat{G}$  are the poles of the components  $\hat{G}_{ij}(s)$ .

*Remark 3.35* As with the SISO case, these definitions are *identical* in discrete and continuous-time. All that changes is a swap of variable from  $s$  to  $z$ .

It's critical to note that the existence of a transmission zero at a point  $s_0 \in \mathbb{C}$  is *not* equivalent to the existence of a vector  $u_0 \neq 0$  for which  $\hat{G}(s_0)u_0 = 0$ ! This is only the case when  $\hat{G}$  has full column normal rank. If this full column rank condition is not met, a point  $s_0 \in \Omega$  can be such that  $\exists u_0 : \hat{G}(s_0)u_0 = 0$  without being a transmission zero.

*Example 3.3 (Transmission Zeros)* Consider the real-rational transfer matrix,

$$\hat{G}(s) = \frac{1}{s+1} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}. \quad (3.217)$$

Here, the normal rank of  $\hat{G}$  is 1, which is less than full column rank. Since there is no  $s$  in the domain of  $\hat{G}$  at which the normal rank *drops*, we conclude that there are no transmission zeros. However, for all  $s$  in the domain of  $\hat{G}$ , there exists a vector  $u_0$  (for example  $u_0 = (1, -1)$ ) for which  $\hat{G}(s)u_0 = 0$ . Thus, it makes sense that we characterize transmission zeros from a drop in rank from the normal rank, rather than from the existence of a vector in the null space of  $\hat{G}(s)$ .

Now that we've come up with a basic definition of the poles and zeros of transfer functions, let's try to gain some intuition for what they actually represent. First, we'll come up with an alternate characterization of poles and zeros.

This lemma provides a basic characterization of poles and zeros in terms of the limiting behavior of the transfer function,  $\hat{G}$ . What we still lack, however, is a time-domain intuition for poles and zeros. First, we'll discuss the connection between the poles of a transfer function and the *eigenvalues* of the  $A$  matrix of a linear system representation.

**Lemma 3.7 (All Poles are Eigenvalues of  $A$ )** Consider a linear, time-invariant system representation  $(A, B, C, D)$ . Every pole of the transfer function  $\hat{G}$  is an eigenvalue of  $A$ .

**Proof** We'll prove this result in the continuous-time case—the discrete-time case is identical, and simply involves trading an  $s$  for a  $z$ . Recall that the transfer function of a linear, time-invariant system representation is computed,

$$\hat{G}(s) = C(sI - A)^{-1}B + D. \quad (3.218)$$



In order for a point  $s_0 \in \mathbb{C}$  to be a pole of  $\hat{G}$ , an entry of  $\hat{G}$  must go to  $\infty$  as  $s \rightarrow s_0$ . Looking at the formula above, the *only way* for this to occur is for the inverse  $(sI - A)^{-1}$  to become ill-posed at  $s_0$ . Since  $sI - A$  is non-invertible precisely at the eigenvalues of  $A$ , we conclude that every pole  $s_0$  of  $\hat{G}$  must be an eigenvalue of  $A$ .  $\square$

Thus, we observe that every pole of a transfer function corresponds to an *eigenvalue* of the  $A$  matrix of the system representation. This lends an important time-domain interpretation of the poles of a linear, time-invariant system. Using the formula for the zero-state response of a continuous-time, LTI system,

$$y(t) = \int_0^t C \exp(A(t - \tau)) B u(\tau) d\tau, \quad (3.219)$$

we find that, when a transfer function has a real pole at  $\lambda$ , the zero-state response to a wide class of input signals will contain a term with  $e^{\lambda t}$ . Likewise, when the transfer function has a complex pole  $\lambda = \sigma + j\omega$ , the zero-state response to a wide class of inputs will contain a damped oscillatory term of the form  $e^{\sigma t} \sin(\omega t + \phi)$ . The terms in the zero-state response corresponding to the poles of the transfer function are often referred to as the *modes* of the zero-state response. Analogous conclusions regarding the presence of decaying and oscillating modes can be made in the discrete-time case.

It's extremely important to note that, in general, the inclusion of the set of poles of a transfer function are a strict subset of the set of eigenvalues of  $A$ . The following example provides one instance in which this is the case.

*Example 3.4 (Pole-Zero Cancellation)* Consider the transfer function,

$$\hat{G}(s) = \frac{1}{s^2 + 2s + 3}. \quad (3.220)$$

Using the results from Chapter 2, we determine that a state space system corresponding to this transfer function is,

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u \quad (3.221)$$

$$y = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}. \quad (3.222)$$

Now, consider the second state-space system,

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ -2 & -3 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} u \quad (3.223)$$

$$y = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}. \quad (3.224)$$

Computing the transfer function of this second system, which we denote  $\hat{G}'(s)$ , we find,

$$\hat{G}'(s) = \frac{s + 1}{(s^2 + 2s + 3)(s + 1)} = \frac{1}{s^2 + 2s + 3} = G(s) \quad (3.225)$$

Thus, the two transfer functions have the *same* poles despite corresponding to different  $A$  matrices. In the first system, the poles of the transfer function were precisely the eigenvalues of the  $A$  matrix, while in the second case, the poles of the transfer function were a strict subset of the eigenvalues of the  $A$  matrix.

Recognizing when two state space representations will have the same transfer function is a surprisingly challenging problem! We'll solve this problem in the next chapter, when we discuss controllability and observability.

Recall that, when studying stability in state space, we showed that the eigenvalues of the  $A$  matrix completely determine the stability of the zero equilibrium of the unforced systems  $\dot{x} = Ax$  and  $x[k+1] = Ax[k]$ . Due to the connection between the poles of a transfer function and the eigenvalues of the  $A$  matrix of a system representation, we make the following definition.

**Definition 3.22 (Stable Transfer Function)** Consider a continuous/discrete-time transfer function  $\hat{G} : \Omega \rightarrow \mathbb{C}^{p \times m}$ .

1. Continuous-time: If  $\hat{G}$  is a continuous-time transfer function, it is said to be stable if all of its poles belong to the open left-half plane,

$$\{s \in \mathbb{C} : \operatorname{Re}(s) < 0\}. \quad (3.226)$$

2. Discrete-time: If  $\hat{G}$  is a discrete-time transfer function, it is said to be stable if all of its poles belong to the open unit disk,

$$\{z \in \mathbb{C} : |z| < 1\}. \quad (3.227)$$

Thus, we declare a transfer function to be *stable* if all of its poles satisfy the exponential stability criteria for eigenvalues.

Now, let's turn our attention to the zeros of a transfer function. Whereas a pole at  $s_0$  corresponds to a response mode of the form  $e^{s_0 t}$  that the system will *have*, a zero at  $s_0$  represents an input signal that the system will *ignore*. Consider the following result.

**Lemma 3.8 (Interpretation of Transmission Zeros)** Consider a continuous-time, LTI system representation  $(A, B, C, D)$  with transfer function  $\hat{G}$ . Suppose  $\hat{G}$  has a transmission zero at  $s_0 \in \mathbb{C}$ , and that  $s_0$  is not an eigenvalue of  $A$ . Then:

1. One has  $\operatorname{rank} \begin{bmatrix} A - s_0 I & B \\ C & D \end{bmatrix} < \operatorname{normalrank} \begin{bmatrix} A - sI & B \\ C & D \end{bmatrix}$ .
2. There exists a vector  $u_0 \in \mathbb{R}^m$  for which the response of the system to the input  $u(t) = u_0 e^{s_0 t}$  from initial condition  $x_0 = (s_0 I - A)^{-1} B u_0$  is  $y(t) \equiv 0$ .

*Remark 3.36* Item (1) of this lemma concerns *invariant zeros*. An invariant zero of the system representation  $(A, B, C, D)$  is a point  $s_0 \in \mathbb{C}$  for which

$$\operatorname{rank} \begin{bmatrix} A - s_0 I & B \\ C & D \end{bmatrix} < \operatorname{normalrank} \begin{bmatrix} A - sI & B \\ C & D \end{bmatrix}. \quad (3.228)$$

*Remark 3.37* In item (2), when  $s_0 \in \mathbb{C}$ , the input  $u_0 e^{s_0 t}$  is *complex*. Thanks to linearity, one can work with complex inputs in the same way as real inputs. This can be formally justified by separately computing the response to the real and imaginary parts of a complex input, and adding the real component to  $j$  times the imaginary component to get the total response of the system.

**Proof** First, we prove item (1). Consider an  $s \in \mathbb{C}$  which is not an eigenvalue of  $A$ . For such an  $s$ , we know that  $(A - sI)$  is invertible. We can therefore compute,

$$\begin{bmatrix} A - sI & B \\ C & D \end{bmatrix} \begin{bmatrix} I - (A - sI)^{-1}B \\ 0 & I \end{bmatrix} = \begin{bmatrix} A - sI - (A - sI)(A - sI)^{-1}B + B \\ C & -C(A - sI)^{-1}B + D \end{bmatrix} \quad (3.229)$$

$$= \begin{bmatrix} A - sI & 0 \\ C & \hat{G}(s) \end{bmatrix}. \quad (3.230)$$

Let's analyze the resulting equality,

$$\underbrace{\begin{bmatrix} A - sI & B \\ C & D \end{bmatrix}}_{\hat{M}_1(s)} \underbrace{\begin{bmatrix} I - (A - sI)^{-1}B \\ 0 & I \end{bmatrix}}_{\hat{M}_2(s)} = \underbrace{\begin{bmatrix} A - sI & 0 \\ C & \hat{G}(s) \end{bmatrix}}_{\hat{M}_3(s)}. \quad (3.231)$$

Since  $s_0$  is a transmission zero of  $\hat{G}(s)$  and not an eigenvalue of  $A$ , it follows that  $s_0$  is a transmission zero of  $\hat{M}_3$ . Since  $\hat{M}_2$  is always full rank, we conclude that  $\hat{M}_1$  must also be a transmission zero of  $M_1$ . Item (1) follows.

Now, we prove item (2). Suppose  $s_0 \in \mathbb{C}$  is a zero of the transfer function, and is not an eigenvalue of  $A$ . Let's compute the response of the system to an arbitrary initial condition  $x_0$  and an input  $u(t) = u_0 e^{s_0 t}$ . We have,

$$\hat{Y}(s) = C(sI - A)^{-1}x_0 + C(sI - A)^{-1}Bu_0(s - s_0)^{-1} + Du_0(s - s_0)^{-1} \quad (3.232)$$

$$= C(sI - A)^{-1}(x_0 - (s_0I - A)^{-1}Bu_0) + \hat{G}(s_0)u_0(s - s_0)^{-1}. \quad (3.233)$$

The initial condition  $x_0 = (s_0I - A)^{-1}Bu_0$  eliminates a number of terms in this expression. For this initial condition, we have,

$$\hat{Y}(s) = \hat{G}(s_0)u_0(s - s_0)^{-1}. \quad (3.234)$$

Since  $\hat{G}$  has a transmission zero at  $s_0$ , there exists a  $u_0$  for which  $\hat{G}(s_0)u_0 = 0$ . Picking such a  $u_0$ , it follows that  $\hat{Y}(s) = 0$ , which implies  $y(t) \equiv 0$ .  $\square$

Thus, the existence of a transmission zero at  $s_0$  implies there exists an input  $u(t) = u_0 e^{s_0 t}$  and an initial condition which the system will entirely *ignore*. As a consequence of this result, one often says that the presence of a zero at  $s_0$  means that the system *blocks* or *absorbs* exponential signals at  $s_0$ .

### 3.3.2 Spaces of Signals

Now, we return to our study of norms on signals and systems. Recall that, at the start of this section, we stated that, in the input/output picture of stability, we're interested in finding how systems augment "size" of different signals. In order to precisely assign a notion of "size" to a signal, we introduce a number of norms on the spaces of continuous and discrete-time signals.

As usual, the set of continuous-time signals we'll consider are piecewise-continuous mappings from  $\mathbb{R}$  to  $\mathbb{R}^m$ , and the set of discrete-time signals we'll consider are sequences, mapping

from  $\mathbb{Z}$  to  $\mathbb{R}^m$ . We note—all of the norms we introduce below are instances of the standard  $\ell^p$  and  $L^p$  norms from functional analysis. Here, we simply highlight a couple of cases that are of particular interest in robust control. In the spirit of robust control, we'll refer to the  $\ell^p$  and  $L^p$  norms simply as the  $p$ -norms, with the choice of  $\ell^p$  versus  $L^p$  being dictated by the time set. First, we consider the *1-norm*, which corresponds to the  $\ell^1$  norm in discrete-time and the  $L^1$  norm in continuous-time.

**Definition 3.23 (1-Norm)** The 1-norm of a signal is defined as follows.

1. Continuous-time: the 1-norm of a continuous-time signal  $u(\cdot) \in PC(\mathbb{R}, \mathbb{R}^m)$  is computed,

$$\|u\|_1 := \int_{\mathbb{R}} \|u(t)\|_1 dt, \quad (3.235)$$

where  $\|u(t)\|_1$  is the  $\ell_1$  norm on  $\mathbb{R}^m$ ,  $\|u(t)\|_1 = \sum_{i=1}^m |u_i(t)|$ . The space of signals  $u(\cdot) \in PC(\mathbb{R}, \mathbb{R}^m)$  with finite 1-norm is denoted  $L^1(\mathbb{R}, \mathbb{R}^m)$ .

2. Discrete-time: the 1-norm of a discrete-time signal  $u[\cdot] : \mathbb{Z} \rightarrow \mathbb{R}^m$  is computed,

$$\|u\|_1 = \sum_{k \in \mathbb{Z}} \|u[k]\|_1, \quad (3.236)$$

where  $\|u[k]\|_1$  is the  $\ell_1$  norm on  $\mathbb{R}^m$ ,  $\|u[k]\|_1 = \sum_{i=1}^m |u_i[k]|$ . The space of signals  $u[\cdot] : \mathbb{Z} \rightarrow \mathbb{R}^m$  with finite 1-norm is denoted  $\ell_1(\mathbb{Z}, \mathbb{R}^m)$ .

*Remark 3.38* When the domain and codomain of the signal are clear, it's convention to suppress the  $(\mathbb{R}, \mathbb{R}^m)$  and  $(\mathbb{Z}, \mathbb{R}^m)$  arguments, and simply write  $L^1$  and  $\ell^1$ .

*Remark 3.39* Formally, the continuous-time 1-norm is not a norm, but a *pseudonorm*. A pseudonorm satisfies every property of a norm with the exception that it is only positive semidefinite instead of positive definite—i.e. it is always the case that  $\|u\|_1 \geq 0$  but it is not necessarily true that  $\|u\|_1 = 0 \Leftrightarrow u(t) \equiv 0$ . This is because the  $L^1$  “norm” is defined in terms of an *integral*. Since integrals neglect the values of functions on sets of measure zero, it could be possible that  $\|u\|_1 = 0$  and  $u(t) \neq 0$  for some  $t \in \mathbb{R}$ . For instance, the scalar signal  $u(t)$  which jumps up to 1 at  $t = 0$  and is zero elsewhere has  $\|u\|_1 = 0$ , but is not the zero signal. One can formally make  $L^1(\mathbb{R}, \mathbb{R}^m)$  into a normed vector space by defining the elements of  $L^1$  not to be signals, but rather sets of signals that are equal everywhere but on sets of measure zero. As this extra layer of abstraction is unnecessary for our treatment of the material, we will proceed with treating elements of  $L^1$  simply as functions. As we won't focus on the details of measure theory in this course, this simplification won't cause us any headaches.

Here, it's critical that we distinguish between the actual signal  $u$  and its value at a particular time,  $u(t)$ . Whenever we write  $u$  without the  $(t)$  argument, we are referring to the *signal*—the actual function from  $\mathbb{R} \rightarrow \mathbb{R}^m$ . Whenever we write  $u(t)$ , we are referring to the *value* of the signal at time  $t$ , which is a vector in  $\mathbb{R}^m$ . Hence,  $\|u\|_1$  refers to the 1-norm of the *signal*, while  $\|u(t)\|_1$  refers to the  $\ell_1$  norm of the *vector*  $u(t)$  at time  $t$ .

To understand what the 1-norm corresponds to, we instantiate the definition of the 1-norm for  $m = 1$ . For a continuous-time signal  $u(\cdot) \in PC(\mathbb{R}, \mathbb{R})$ , one has

$$\|u\|_1 = \int_{\mathbb{R}} |u(t)| dt, \quad (3.237)$$

that the 1-norm is simply the integral of the magnitude of  $u(t)$  over  $\mathbb{R}$ . Likewise, in the discrete-time case, for  $u : \mathbb{Z} \rightarrow \mathbb{R}$  one has

$$\|u\|_1 = \sum_{k \in \mathbb{Z}} |u[k]|. \quad (3.238)$$

Also of interest to us is the 2-norm of a signal.

**Definition 3.24 (2-Norm)** The 2-norm of a signal is defined as follows.

1. Continuous-time: the 2-norm of a continuous-time signal  $u(\cdot) \in PC(\mathbb{R}, \mathbb{R}^m)$  is computed,

$$\|u\|_2 := \sqrt{\int_{\mathbb{R}} \|u(t)\|_2^2 dt}, \quad (3.239)$$

where  $\|u(t)\|_2$  is the  $\ell_2$  norm on  $\mathbb{R}^m$ ,  $\|u(t)\|_2 = \sqrt{\sum_{i=1}^m (u_i(t))^2}$ . The space of signals  $u(\cdot) \in PC(\mathbb{R}, \mathbb{R}^m)$  with finite 2-norm is denoted  $L^2(\mathbb{R}, \mathbb{R}^m)$ .

2. Discrete-time: the 2-norm of a discrete-time signal  $u[\cdot] : \mathbb{Z} \rightarrow \mathbb{R}^m$  is computed,

$$\|u\|_2 = \sqrt{\sum_{k \in \mathbb{Z}} \|u[k]\|_2^2}, \quad (3.240)$$

where  $\|u[k]\|_2$  is the  $\ell_2$  norm on  $\mathbb{R}^m$ ,  $\|u[k]\|_2 = \sqrt{\sum_{i=1}^m (u_i[k])^2}$ . The space of signals  $u[\cdot] : \mathbb{Z} \rightarrow \mathbb{R}^m$  with finite 2-norm is denoted  $\ell^2(\mathbb{Z}, \mathbb{R}^m)$ .

*Remark 3.40* As in the case of the continuous-time 1-norm, the continuous-time 2-norm is formally a *pseudonorm*, not a norm. Once again, the distinction between  $L_2$  norm and  $L_2$  pseudonorm will have very minimal consequences for our analysis in this course.

In the scalar case, the 2-norm of a continuous-time signal  $u(\cdot) \in PC(\mathbb{R}, \mathbb{R})$  is computed,

$$\|u\|_2 = \sqrt{\int_{\mathbb{R}} (u(t))^2 dt}, \quad (3.241)$$

as the integral of the square of  $u(t)$  over  $\mathbb{R}$ . Likewise, the 2-norm of a discrete-time signal  $u[\cdot] : \mathbb{Z} \rightarrow \mathbb{R}$  is computed,

$$\|u\|_2 = \sqrt{\sum_{k \in \mathbb{Z}} (u[k])^2}. \quad (3.242)$$

The square of the 2-norm of a signal is often interpreted as the *energy* of the signal. This is due to connections between the 2-norm and energy dissipated by a resistor in an electrical circuit. For an ideal resistor with resistance  $R$  and time-varying voltage  $v(t)$ , the energy dissipated by the resistor across all time is computed,

$$E = \frac{1}{R} \int_{\mathbb{R}} (v(t))^2 dt. \quad (3.243)$$

For a resistance of  $R = 1$  Ohm,  $E$  coincides with the square of the 2-norm of  $v(\cdot)$ .<sup>4</sup> The final signal norm we'll discuss is the  $\infty$ -norm.

**Definition 3.25 ( $\infty$ -Norm)** The  $\infty$ -norm of a signal is defined as follows.

1. Continuous-time: the  $\infty$ -norm of a continuous-time signal  $u(\cdot) \in PC(\mathbb{R}, \mathbb{R}^m)$  is computed,

$$\|u\|_{\infty} := \operatorname{ess\,sup}_{t \in \mathbb{R}} \|u(t)\|_{\infty}, \quad (3.244)$$

where  $\|u(t)\|_{\infty}$  is the  $\ell_{\infty}$  norm on  $\mathbb{R}^m$ ,  $\|u(t)\|_{\infty} = \max_{i=1, \dots, m} |u_i(t)|$ . The space of signals  $u(\cdot) \in PC(\mathbb{R}, \mathbb{R}^m)$  with finite  $\infty$ -norm is denoted  $L^{\infty}(\mathbb{R}, \mathbb{R}^m)$ .

2. Discrete-time: the  $\infty$ -norm of a discrete-time signal  $u[\cdot] : \mathbb{Z} \rightarrow \mathbb{R}^m$  is computed,

$$\|u\|_{\infty} = \sup_{k \in \mathbb{Z}} \|u[k]\|_{\infty}, \quad (3.245)$$

where  $\|u[k]\|_{\infty}$  is the  $\ell_{\infty}$  norm on  $\mathbb{R}^m$ . The space of signals  $u : \mathbb{Z} \rightarrow \mathbb{R}^m$  with finite  $\infty$ -norm is denoted  $\ell_{\infty}^m$ .

*Remark 3.41* The notation  $\operatorname{ess\,sup}_{t \in \mathbb{R}} \|u(t)\|_{\infty}$  refers to the *essential supremum* of  $\|u(t)\|_{\infty}$ . As opposed to the standard supremum, which must be an upper bound for all  $t \in \mathbb{R}$ , the essential supremum simply needs to be an upper bound for *almost every*  $t \in \mathbb{R}$  (where *almost every*  $t \in \mathbb{R}$  means for every  $t \in \mathbb{R}$  except for a set of measure zero). For a function  $f : \mathbb{R} \rightarrow \mathbb{R}$ , the essential supremum  $\operatorname{ess\,sup}_{t \in \mathbb{R}} f(t)$  is defined,

$$\inf\{C \in \mathbb{R} : f(t) \leq C \text{ for almost every } t \in \mathbb{R}\}. \quad (3.246)$$

Thus,  $\operatorname{ess\,sup}_{t \in \mathbb{R}} f(t)$  is the smallest value of  $C$  for which  $f(t) \leq C$  for all  $t \in \mathbb{R}$  except for a set of measure zero. Since integrals neglect the values of signals on sets of measure zero, the essential supremum is appropriate for the study of continuous-time signals. As with the case of the 1 and 2-norms, the continuous-time  $\infty$ -norm is a pseudonorm. This is due to the essential supremum being an upper bound for almost every  $t \in \mathbb{R}$  instead of every  $t$ .

The  $\infty$ -norm is often interpreted as the norm which gives the *worst-case* value of a signal. This is because it scans over *all time* and *all entries* to find an upper bound on the size of the signal that holds for (almost all)  $t$ . These three norms—the 1-norm, 2-norm, and  $\infty$ -norm, are of primary interest to us in the I/O study of stability.

### 3.3.3 Spaces of Systems

With a set of norms of interest on *signals*, we now return to the problem of determining the norms on *signals*. At the start of the section, we mentioned that - given a norm  $\|\cdot\|$  on a signal - a natural choice of norm on the *system* would be the smallest constant  $C$  for which

$$\|G * u\| \leq C \|u\|, \quad (3.247)$$

---

<sup>4</sup> Thanks to Yiheng Xie for providing this example!

for all admissible inputs  $u$  to the system. Thus, what we'd like to find is the *induced operator norm* of  $G*$ , the operator of convolution with the impulse response,  $G$ . This leads us to the following definition.

**Definition 3.26 (Induced System Norm)** Consider a linear, time-invariant system with impulse response map  $G$ . Let  $p, q \in \{1, 2, \infty\}$ . Suppose the input signal space has norm  $\|\cdot\|_p$ , and the output signal space has norm  $\|\cdot\|_q$ . The induced  $p$  to  $q$  system norm is defined,

$$\|G\|_{p,q} := \sup_{u \neq 0} \frac{\|G * u\|_q}{\|u\|_p} = \sup_{\|u\|_p=1} \|G * u\|_q. \quad (3.248)$$

*Remark 3.42* The notation  $\|G\|_{p,q}$  seems to suggest that  $\|G\|_{p,q}$  is the norm of the impulse response signal. However, this is not the case! It's important to remember that  $\|G\|_{p,q}$  represents the induced operator norm of  $G*$ , the operator of convolution with  $G$ .

*Remark 3.43* The induced system norm from  $p$  to  $q$  is also referred to as the  $p$ - $q$  gain of  $G$ .

We now confirm that the induced system norm  $\|G\|_{p,q}$  is in fact the smallest constant  $C \geq 0$  for which  $\|G * u\|_q \leq C \|u\|_p$ , for all admissible inputs  $u$  with  $\|u\|_p < \infty$ .

**Proposition 3.10 (Induced Norm is the Smallest System Gain)** Let  $G$  be an impulse response map between signal spaces with norms  $\|\cdot\|_p$  and  $\|\cdot\|_q$ . The induced  $p$  to  $q$  system norm of  $G$  is equivalently calculated,

$$\|G\|_{p,q} = \inf\{C \geq 0 : \|G * u\|_q \leq C \|u\|_p, \forall u \in \mathcal{U} \text{ s.t. } \|u\|_p < \infty\}. \quad (3.249)$$

**Proof** First, we note that, by linearity of convolution,

$$\inf\{C \geq 0 : \|G * u\|_q \leq C \|u\|_p, \forall u \in \mathcal{U} \text{ s.t. } \|u\|_p < \infty\} \quad (3.250)$$

$$= \inf\{C \geq 0 : \|G * u\|_q \leq C, \forall u \in \mathcal{U} \text{ s.t. } \|u\|_p = 1\}. \quad (3.251)$$

Thus, we have that the proposed value is the least upper bound on  $\{\|G * u\|_q : \|u\|_p \leq 1\}$ . Since this is the definition of the supremum, we conclude,

$$\inf\{C \geq 0 : \|G * u\|_q \leq C, \forall u \in \mathcal{U} \text{ s.t. } \|u\|_p = 1\} = \sup_{\|u\|_p=1} \|G * u\|_q = \|G\|_{p,q}. \quad (3.252)$$

We conclude that the proposed value equals the induced system norm.  $\square$

Using the induced system norm, we can formulate a precise definition of input/output stability. The basic idea is the following: if, for every input  $u$  of finite  $p$  norm, the output is of finite  $q$  norm, the system is  $p$  to  $q$  stable. Otherwise, if there exists an input  $u$  of finite  $p$  norm for which the output has infinite  $q$  norm, then the system is *not*  $p$  to  $q$  stable.

**Definition 3.27 (p-q Stability)** Consider a linear, time-invariant system with impulse response map  $G$ . Let  $p, q \in \{1, 2, \infty\}$ . The system is  $p$ - $q$  stable if  $\|G\|_{p,q} < \infty$ .

Why does this make sense as a definition of input to output stability? Intuitively, a system *should* be input to output stable if the output of the system doesn't "blow up" for an input signal of finite size. This is *all* that the definition of  $p$ - $q$  stability is saying—there is *no* input of finite  $p$  norm which will cause the output to be of infinite  $q$ -norm. A case of particular interest is when  $p, q = \infty$ .

**Definition 3.28 (BIBO Stability)** A linear, time-invariant system with impulse response map  $G$  is bounded input, bounded output (BIBO) stable if  $\|G\|_{\infty, \infty} < \infty$ .

The name bounded input, bounded output stability derives directly from the definition of the  $\infty$  norm,

$$\|u\|_{\infty} = \operatorname{ess\,sup}_{t \in \mathbb{R}} \|u(t)\|_{\infty} \quad (\text{Continuous-time}) \quad (3.253)$$

$$\|u\|_{\infty} = \sup_{k \in \mathbb{Z}} \|u[k]\|_{\infty} \quad (\text{Discrete-time}). \quad (3.254)$$

As can be seen from the definitions of the continuous and discrete-time  $\infty$ -norms, a signal is of bounded  $\infty$ -norm if and only if it is (essentially) bounded across its domain. Thus, a system which is bounded input, bounded output stable ensures that every (essentially) bounded input signal results in an (essentially) bounded output signal.

Now that we've introduced definitions for induced system norms and input to output stability, we're faced with an important question. How do we actually *compute* the norms on our systems? To answer this question, we go back to the key idea that we introduced at the beginning of the section: that *time domain* performance specifications manifest themselves in *frequency domain* conditions.

In order to actually calculate the  $p$  to  $q$  norms of a system, we'll find it fruitful to first define norms on *transfer functions*. Once we've defined a suitably rich set of norms on transfer functions, we'll be able to derive expressions for each of the  $p$  to  $q$  system norms. Thus, the remainder of the section will proceed as follows. First, we'll define two important norms on the space of transfer functions. Then, we'll study the response of transfer functions to a few important signals. Finally, we'll use the response of the systems to these signals to compute a table of  $p$  to  $q$  norms.

### 3.3.3.1 Norms on Transfer Functions

As alluded to above, in order to calculate the  $p$  to  $q$  norm of a system, we'll need to understand how to place a norm on the system's *transfer function*. The first step in constructing a norm on the transfer function is recognizing that the response of the system to oscillatory inputs reveals a great deal about its behavior. In the following proposition, we compute the steady state response of a system to an oscillatory input.

**Proposition 3.11 (Frequency Response)** Let  $\hat{G}$  be a stable, SISO transfer function.

1. *Continuous-time:* suppose  $\hat{G}$  is a continuous-time transfer function with no pole at  $s = j\omega$ . As  $t \rightarrow \infty$ , the zero-state response of the system to the complex input  $u(\cdot) : \mathbb{R}_{\geq 0} \rightarrow \mathbb{C}$ ,  $u(t) = e^{j\omega t}$  converges pointwise to the signal,

$$y(t) = \hat{G}(j\omega)e^{j\omega t}. \quad (3.255)$$

2. *Discrete-Time:* suppose  $\hat{G}$  is a discrete-time transfer function with no pole at  $z = j\omega$ . As  $k \rightarrow \infty$ , the zero-state response of the system to the complex input  $u[\cdot] : \mathbb{Z}_{\geq 0} \rightarrow \mathbb{C}$ ,  $u[k] = e^{j\omega k}$  converges pointwise to the signal,

$$y(t) = \hat{G}(e^{j\omega})e^{j\omega k}. \quad (3.256)$$



*Remark 3.44* Recall that  $e^{j\omega t} = \cos(\omega t) + j \sin(\omega t)$ . The signal  $e^{j\omega t}$  is therefore a complex, oscillatory input defined by sines and cosines of a frequency  $\omega$ .

**Proof** See Problem 3.20 for the details.  $\square$

Thus, we find that, for a stable transfer function, the size of the steady state response to an oscillatory signal is governed by the size of  $\hat{G}(j\omega)$  (in continuous-time) and  $\hat{G}(e^{j\omega})$  in discrete-time. This result leads us to make the following definition.

**Definition 3.29 (Frequency Response)** Consider a transfer function  $\hat{G} : \Omega \rightarrow \mathbb{C}^{p \times m}$ .

1. Continuous-time: suppose  $\hat{G} : \Omega \rightarrow \mathbb{C}^{p \times m}$  is a continuous-time transfer function. The frequency response of  $\hat{G}$  is the map  $\hat{G}_\omega : \Omega_{j\mathbb{R}} \rightarrow \mathbb{C}^{p \times m}$ , defined

$$\hat{G}_\omega(\omega) = \hat{G}(j\omega), \quad \forall \omega \in \Omega_{j\mathbb{R}} = \{\omega \in \mathbb{R} : j\omega \in \Omega\}. \quad (3.257)$$

2. Discrete-time: suppose  $\hat{G} : \Omega \rightarrow \mathbb{C}^{p \times m}$  is a discrete-time transfer function. The frequency response of  $\hat{G}$  is the map  $\hat{G}_\omega : \Omega_{e^{j\omega}} \rightarrow \mathbb{C}^{p \times m}$ , defined

$$\hat{G}_\omega(\omega) = \hat{G}(e^{j\omega}), \quad \forall \omega \in \Omega_{e^{j\omega}} = \{\omega \in [-\pi, \pi) : e^{j\omega} \in \Omega\}. \quad (3.258)$$

*Remark 3.45* Note that one typically refers to  $\hat{G}(j\omega)$  and  $\hat{G}(e^{j\omega})$  as the frequency response, rather than the actual map  $\hat{G}_\omega(\omega)$ . For instance, one might encounter a line such as, “consider the frequency response  $\hat{G}(j\omega)$ .”

Thus, the frequency response of a given transfer function is simply the value of the transfer function which determines the steady state response of the system to an oscillatory input of frequency  $\omega$ . As it happens, the class of oscillatory inputs  $e^{j\omega t}$  and  $e^{j\omega k}$  are so revealing for LTI systems that the frequency responses  $\hat{G}(j\omega)$  and  $\hat{G}(e^{j\omega})$  can be used to define norms on the transfer functions  $\hat{G}(s)$  and  $\hat{G}(z)$ . In the following definition, we translate some familiar features of the time domain  $L^p$  norms to the frequency domain in order to define norms on transfer functions.

**Definition 3.30 ( $\mathcal{H}_2, \mathcal{H}_\infty$  Transfer Function Norms)** Consider a matrix-valued transfer function  $\hat{G}$ . The  $\mathcal{H}_2$  and  $\mathcal{H}_\infty$  norms of  $\hat{G}$  are defined as follows.

1. Continuous-time  $\mathcal{H}_2$ -norm: the continuous-time  $\mathcal{H}_2$ -norm of  $\hat{G}$  is defined,

$$\|\hat{G}\|_2 := \sqrt{\frac{1}{2\pi} \int_{\mathbb{R}} \text{tr}(\hat{G}^*(j\omega) \hat{G}(j\omega)) d\omega}. \quad (3.259)$$

2. Discrete-time  $\mathcal{H}_2$ -norm: the discrete-time  $\mathcal{H}_2$ -norm of  $\hat{G}$  is defined,

$$\|\hat{G}\|_2 := \sqrt{\frac{1}{2\pi} \int_{-\pi}^{\pi} \text{tr}(\hat{G}^*(e^{j\omega}) \hat{G}(e^{j\omega})) d\omega}. \quad (3.260)$$

3. Continuous-time  $\mathcal{H}_\infty$ -norm: the continuous-time  $\mathcal{H}_\infty$ -norm of  $\hat{G}$  is defined,

$$\|\hat{G}\|_\infty := \sup_{\omega \in \mathbb{R}} \sigma_{\max}(\hat{G}(j\omega)). \quad (3.261)$$

4. Discrete-time  $\mathcal{H}_\infty$ -norm: the discrete-time  $\mathcal{H}_\infty$ -norm of  $\hat{G}$  is defined,

$$\|\hat{G}\|_\infty := \sup_{\omega \in [-\pi, \pi)} \sigma_{\max}(\hat{G}(e^{j\omega})). \quad (3.262)$$

*Remark 3.46* Our definitions of the  $\mathcal{H}_p$ -norms are consistent with the those used in robust control, but are *slightly* different to those used in mathematics. The definition posed above coincides with the definition of the  $\mathcal{H}_p$ -norm found in mathematics for the class of *subharmonic functions* [7].

*Remark 3.47* Unfortunately, the conventions of the field dictate that we have some notational overload on our hands. One must be careful to distinguish between,

$$\|G(t)\|_2, \|G\|_2, \text{ and } \|\hat{G}\|_2 \quad (3.263)$$

$$\|G(t)\|_\infty, \|G\|_\infty, \text{ and } \|\hat{G}\|_\infty. \quad (3.264)$$

Whenever we write  $(t)$ , we mean that we are taking the norm of a single vector or matrix value. When we don't write  $(t)$  and don't write a hat, we mean that we are taking the time domain signal norm (e.g. 2-norm,  $\infty$ -norm). Finally, when we write a hat, we mean that we are taking the frequency domain norm (e.g.  $\mathcal{H}_2$ -norm,  $\mathcal{H}_\infty$ -norm) of a transfer function. If you're ever confused about which norm is being used, check for these elements! The  $(t)$ , no  $(t)$  and no hat, and hat will always dictate which norm is in use.

*Remark 3.48* Notice the similarity in structure between the  $\mathcal{H}_2$  transfer function norm and the  $L^2$  signal norm! In the  $\mathcal{H}_2$  norm, we integrate a “square” term (the term  $\text{tr}(\hat{G}^*(j\omega)\hat{G}(j\omega))$ ), just like how we integrate the square of the Euclidean 2-norm when computing the  $L^2$  signal norm. Likewise, there exist parallels between the  $\mathcal{H}_\infty$  transfer function norm and the  $L^\infty$  signal norm. In the  $\mathcal{H}_\infty$  norm, we take a supremum over frequency, similar to  $L^\infty$  norm case, where we take an (essential) supremum over time.

As is the case of the  $p$ -norms for signals, the  $\mathcal{H}_p$  norms for transfer functions take on a dramatically simpler form in the event where one has a scalar transfer function (corresponding to a SISO system). In this case, the  $\mathcal{H}_p$ -norms reduce to,

$$\|\hat{G}\|_2 = \sqrt{\frac{1}{2\pi} \int_{\mathbb{R}} |\hat{G}(j\omega)|^2 d\omega} \quad (\text{Continuous-time } \mathcal{H}_2\text{-norm}) \quad (3.265)$$

$$\|\hat{G}\|_2 = \sqrt{\frac{1}{2\pi} \int_{-\pi}^{\pi} |\hat{G}(e^{j\omega})|^2 d\omega} \quad (\text{Discrete-time } \mathcal{H}_2\text{-norm}) \quad (3.266)$$

$$\|\hat{G}\|_\infty = \sup_{\omega \in \mathbb{R}} |\hat{G}(j\omega)| \quad (\text{Continuous-time } \mathcal{H}_\infty\text{-norm}) \quad (3.267)$$

$$\|\hat{G}\|_\infty = \sup_{\omega \in [-\pi, \pi)} |\hat{G}(e^{j\omega})| \quad (\text{Discrete-time } \mathcal{H}_\infty\text{-norm}). \quad (3.268)$$

Just like we defined the  $L^p$  and  $\ell^p$  spaces based on signals having finite  $L^p$  and  $\ell^p$  norms, we can define spaces of *transfer functions* based on whether or not they have finite  $\mathcal{H}_p$ -norms.

**Definition 3.31 ( $\mathcal{RH}_2$ ,  $\mathcal{RH}_\infty$  Spaces)** The continuous and discrete-time  $\mathcal{RH}_2$  and  $\mathcal{RH}_\infty$  spaces of transfer functions are defined as follows.

1. Continuous-time  $\mathcal{RH}_2$  space: the continuous-time  $\mathcal{RH}_2$  space consists of all proper, real-rational matrix transfer functions that have finite (continuous-time)  $\mathcal{H}_2$ -norm and are analytic in the open right-half plane,  $\{s \in \mathbb{C} : \operatorname{Re}(s) > 0\}$ .
2. Discrete-time  $\mathcal{RH}_2$  space: the discrete-time  $\mathcal{RH}_2$  space consists of all proper, real-rational matrix transfer functions that have finite (discrete-time)  $\mathcal{H}_2$ -norm and are analytic outside the unit disk,  $\{z \in \mathbb{C} : |z| > 1\}$ .
3. Continuous-time  $\mathcal{RH}_\infty$  space: the continuous-time  $\mathcal{RH}_\infty$  space consists of all proper, real-rational matrix transfer functions that have finite (continuous-time)  $\mathcal{H}_\infty$ -norm and are analytic in the open right-half plane,  $\{s \in \mathbb{C} : \operatorname{Re}(s) > 0\}$ .
4. Discrete-time  $\mathcal{RH}_\infty$  space: the discrete-time  $\mathcal{RH}_\infty$  space consists of all proper, real-rational matrix transfer functions that have finite (discrete-time)  $\mathcal{H}_\infty$ -norm and are analytic outside the unit disk,  $\{z \in \mathbb{C} : |z| > 1\}$ .

*Remark 3.49* Recall that an analytic function is a function that is *complex differentiable*. For rational functions, it's a fairly simple task to determine whether a function is analytic at a given point. Since rational functions are the ratios of polynomials, one may verify that a rational function is analytic at a point if its denominator polynomial has no root there. Shortly, we'll make use of this fact to determine an alternate characterization of the  $\mathcal{RH}_2$  and  $\mathcal{RH}_\infty$  spaces in terms of stability.

The spaces  $\mathcal{RH}_2$  and  $\mathcal{RH}_\infty$  are subsets of more general classes of complex matrix functions called *Hardy spaces*—these spaces are what the  $\mathcal{H}$  in  $\mathcal{RH}_2$  and  $\mathcal{RH}_\infty$  refers to. The  $\mathcal{R}$  in each of the names of course refers to the *rationality* of the transfer functions.

Although the definition posed above lends itself well to the full, general definition of a Hardy space, it's perhaps not the most efficient definition for our purposes. What we'd like is a quick and easy way to determine if a given transfer function is in  $\mathcal{RH}_2$  or  $\mathcal{RH}_\infty$ . Fortunately, we can come up with a precise, equivalent characterization of both spaces in terms of *properness* and *stability*.

**Theorem 3.12 (Equivalent Characterizations of  $\mathcal{RH}_p$ )** Let  $\hat{G} : \Omega \rightarrow \mathbb{C}^{p \times m}$  be a real rational transfer function of a continuous-time system.

1. Continuous-time  $\mathcal{RH}_2$  space:  $\hat{G} \in \mathcal{RH}_2$  if and only if it is strictly proper and stable.
2. Discrete-time  $\mathcal{RH}_2$  space:  $\hat{G} \in \mathcal{RH}_2$  if and only if it is proper and stable.
3. Continuous-time  $\mathcal{RH}_\infty$  space:  $\hat{G} \in \mathcal{RH}_\infty$  if and only if it is proper and stable.
4. Discrete-time  $\mathcal{RH}_\infty$  space:  $\hat{G} \in \mathcal{RH}_\infty$  if and only if it is proper and stable.

**Exercise 3.25** Convince yourself of the claim of Theorem 3.12.

With this simple characterization of  $\mathcal{RH}_2$  and  $\mathcal{RH}_\infty$  in hand, it becomes easy to show that  $\mathcal{RH}_2$  and  $\mathcal{RH}_\infty$  are *vector spaces*. Somewhat harder to prove is that, under the  $\mathcal{H}_2$  and  $\mathcal{H}_\infty$  norms,  $\mathcal{RH}_2$  and  $\mathcal{RH}_\infty$  are *normed vector spaces*.

**Theorem 3.13 ( $\mathcal{RH}_p$  is a Normed Vector Space)** Let  $p \in \{2, \infty\}$ . When equipped with the continuous/discrete-time  $\mathcal{H}_p$ -norm, the continuous/discrete-time  $\mathcal{RH}_p$  space is a normed vector space over  $\mathbb{R}$ .

It's important to note that the spaces  $\mathcal{RH}_2$  and  $\mathcal{RH}_\infty$  are *not* complete normed vector spaces—one can construct convergent sequences of rational functions that escape these spaces. However, one may show that the more general class of Hardy spaces *are* in fact

complete normed vector spaces. What's more, the Hardy space  $\mathcal{H}_2$ , which subsumes  $\mathcal{RH}_2$ , has the structure of a *Hilbert space*—a complete normed vector space in which the norm is induced by an inner product.

**Exercise 3.26** Argue that the  $\mathcal{H}_2$ -norm is induced by an inner product.

### 3.3.3.2 Computing $p$ to $q$ Norms

So far, we've introduced a definition for  $p$ - $q$  stability based on the induced operator norm  $\|G\|_{p,q}$  of convolution with an impulse response map. We established that, in order to calculate the induced operator norm, we must calculate the supremum,

$$\sup_{\|u\|_p=1} \|G * u\|_q. \quad (3.269)$$

To make inroads on this problem, we took a detour to define norms on the space of transfer functions. With these transfer function norms in hand, we're ready to return to the problem of computing  $\|G\|_{p,q}$ .

There are a couple of useful techniques for computing induced system norms. One method of computing the system norm  $\|G\|_{p,q}$  is to exhibit a *sharp* upper bound on  $\|G * u\|_q$  across all admissible input signals  $u$  satisfying  $\|u\|_p = 1$ . In this method, one identifies an upper bound on  $\|G * u\|_q$  across all inputs with  $\|u\|_p = 1$ , and then exhibits an input  $u$  with  $\|u\|_p = 1$  that achieves the upper bound. If an upper bound is achieved, it *must* equal the supremum—hence, this method lets us compute the system norm. In order to employ this “sharp upper bound” method, it's be helpful to know the norm  $\|G * u\|_q$  for a few basic input signals. In the following lemma, we present the norm of  $G * u$  for two inputs: an impulse function and a sine function.

**Lemma 3.9 (Response of an LTI System to Signals)** *Consider a MIMO, LTI system with an impulse response map  $G$  and a transfer function  $\hat{G} \in \mathcal{RH}_\infty$ .*

1. *Continuous-time: suppose the system is continuous-time, with impulse response map  $G(\cdot) : \mathbb{R} \rightarrow \mathbb{R}^{p \times m}$ . Fix a vector  $u_0 \in \mathbb{R}^m$  and a constant  $\omega > 0$ . The response of the system to the inputs  $u(t) = u_0 \delta(t)$  and  $u(t) = u_0 \sin(\omega t)$  satisfies the table,*

| Output/Input   | $u(t) = u_0 \delta(t)$ | $u(t) = u_0 \sin(\omega t)$                |
|----------------|------------------------|--|
| $\ y\ _2$      | $\ Gu_0\ _2$           | $\infty$ if $\hat{G}(j\omega)u_0 \neq 0$   |
| $\ y\ _\infty$ | $\ Gu_0\ _\infty$      | $\geq \max_i  \hat{G}_i^\top(j\omega)u_0 $ |

(3.270)

where  $\hat{G}_i^\top$  refers to the  $i$ 'th row of  $\hat{G}$ .

2. *Discrete-time: suppose the system is discrete-time, with impulse response map  $G[\cdot] : \mathbb{Z} \rightarrow \mathbb{R}^{p \times m}$ . Fix a vector  $u_0 \in \mathbb{R}^m$  and a constant  $\omega > 0$ . The response of the system to the inputs  $u[k] = u_0 \delta[k]$  and  $u[k] = u_0 \sin(\omega k)$  satisfies the table,*

|                |                       |  |
|----------------|-----------------------|--|
| Output/Input   | $u[k] = u_0\delta[k]$ | $u[k] = u_0 \sin(\omega k)$                    |
| $\ y\ _2$      | $\ Gu_0\ _2$          | $\infty$ if $\hat{G}(e^{j\omega})u_0 \neq 0$   |
| $\ y\ _\infty$ | $\ Gu_0\ _\infty$     | $\geq \max_i  \hat{G}_i^\top(e^{j\omega})u_0 $ |

(3.271)

where  $\hat{G}_i^\top$  refers to the  $i$ 'th row of  $\hat{G}$ .

**Remark 3.50** Notice that for each fixed vector  $u_0 \in \mathbb{R}^m$ ,  $G(\cdot)u_0$  is a vector-valued signal! We can therefore apply the standard signal norms we discussed above to  $G(\cdot)u_0$ .

**Proof** I will add a detailed proof for this result a little later in the quarter! In the meantime, you can refer to [43] for a proof of the full continuous-time MIMO case, and [12] for a more elementary proof of the continuous-time SISO case. You can find proofs of the discrete-time case in [10].  $\square$

What conclusions can we draw from this basic table of norms? Importantly, we find that, even though every norm we take is in the time domain, frequency domain quantities appear! Thus, we observe that the frequency domain description of the system is fundamentally embedded in the problem of computing the system's norm. Inspired by this observation, we now seek a way to connect the frequency domain norms on transfer functions with the time domain quantities we'd like to compute. *Parseval's theorem* enables us to do this.

**Theorem 3.14 (Parseval's Theorem)** *The following hold:*

1. *Continuous-time:* consider a continuous-time matrix-valued signal  $G(\cdot) : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^{p \times m}$ , for which  $\int_0^\infty \text{tr}(G^\top(t)G(t))dt < \infty$ . If  $\hat{G}$  denotes the Laplace transform of  $G$ , then

$$\int_{\mathbb{R}_{\geq 0}} \text{tr}(G^\top(t)G(t))dt = \frac{1}{2\pi} \int_{\mathbb{R}} \text{tr}(\hat{G}^*(j\omega)\hat{G}(j\omega))d\omega = \|\hat{G}\|_2^2. \quad (3.272)$$

2. *Discrete-Time:* consider a discrete-time matrix-valued signal  $G[\cdot] : \mathbb{Z}_{\geq 0} \rightarrow \mathbb{R}^{p \times m}$ , for which  $\sum_{k=0}^\infty \text{tr}(G^\top[k]G[k]) < \infty$ . If  $\hat{G}$  denotes the  $\mathcal{Z}$ -transform of  $G$ , then

$$\sum_{k=0}^\infty \text{tr}(G^\top[k]G[k]) = \frac{1}{2\pi} \int_{-\pi}^\pi \text{tr}(\hat{G}^*(e^{j\omega})\hat{G}(e^{j\omega}))d\omega = \|\hat{G}\|_2^2. \quad (3.273)$$

**Remark 3.51** In engineering circles, this result is most commonly known as *Parseval's theorem*, while in mathematics, it is more commonly known as *Plancherel's theorem*. We'll stick with using the name Parseval here.

**Remark 3.52** When specialized from a matrix-valued signal to a vector-valued signal  $u$  and its transform  $\hat{U}$ , Parseval's theorem states that the  $L_2$  norm of  $u$  in the time domain equals the  $\mathcal{H}_2$  norm of its transform  $\hat{U}$ ,  $\|u\|_2 = \|\hat{U}\|_2$ .

**Proof** A rigorous proof of this result requires analysis which takes us somewhat far afield of our course. The details of the scalar case may be found in [34], Chapter 9. This scalar proof generalizes straightforwardly to the matrix case.  $\square$

Armed with Parseval's Theorem, we're finally ready to compute induced system norms.

**Theorem 3.15 (Induced System Norms)** *Consider a MIMO, LTI system with impulse response map  $G$  and transfer function  $\hat{G} \in \mathcal{RH}_\infty$ .*

1. *Continuous-time: suppose the system is continuous-time, with impulse response map  $G(\cdot) : \mathbb{R} \rightarrow \mathbb{R}^{p \times m}$ . The system norms respect the following table:*

| <i>Norm</i>    | $\ u\ _2$  | $\ u\ _\infty$  |
|----------------|--|---|
| $\ y\ _2$      | $\ G\ _{2,2} = \ \hat{G}\ _\infty$<br>$= \sup_{\omega \in \mathbb{R}} \sigma_{\max}(\hat{G}(j\omega))$                                 | $\ G\ _{\infty,2} = \infty$<br><i>(if <math>G</math> is nonzero)</i>  |
| $\ y\ _\infty$ | $\ G\ _{2,\infty} = \ \hat{G}\ _2$<br>$= \sqrt{\frac{1}{2\pi} \int_{\mathbb{R}} \text{tr}(\hat{G}^*(j\omega)\hat{G}(j\omega))d\omega}$ | $\ G\ _{\infty,\infty} = \max_{i=1,\dots,p} \ G_i^\top\ _1$<br>$= \max_i \int_{\mathbb{R}_{\geq 0}} \ G_i^\top(t)\ _1 dt$ |

(3.274)

where  $G_i^\top$  denotes the  $i$ 'th row of  $G$ .

2. *Discrete-time: suppose the system is discrete-time, with impulse response map  $G[\cdot] : \mathbb{Z} \rightarrow \mathbb{R}^{p \times m}$ . The system norms respect the following table:*

| <i>Norm</i>    | $\ u\ _2$  | $\ u\ _\infty$  |
|----------------|--|---|
| $\ y\ _2$      | $\ G\ _{2,2} = \ \hat{G}\ _\infty$<br>$= \sup_{\omega \in [-\pi, \pi)} \sigma_{\max}(\hat{G}(e^{j\omega}))$                                    | $\ G\ _{\infty,2} = \infty$<br><i>(if <math>G</math> is nonzero)</i>  |
| $\ y\ _\infty$ | $\ G\ _{2,\infty} = \ \hat{G}\ _2$<br>$= \sqrt{\frac{1}{2\pi} \int_{-\pi}^{\pi} \text{tr}(\hat{G}^*(e^{j\omega})\hat{G}(e^{j\omega}))d\omega}$ | $\ G\ _{\infty,\infty} = \max_{i=1,\dots,p} \ G_i^\top\ _1$<br>$= \max_i \sum_{k=0}^{\infty} \ G_i^\top[k]\ _1$ |

(3.275)

where  $G_i^\top$  denotes the  $i$ 'th row of  $G$ .

**Remark 3.53** The tables above should be read as follows: each column represents a choice of norm on the *input*, while each row represents a choice of norm on the *output*. Each table entry is the induced system norm corresponding to the choices of norms on the input and output. For instance, the entry  $\|\hat{G}\|_\infty$  belonging to the  $\|u\|_2, \|y\|_2$  entry means that the induced system norm from 2-norm to 2-norm is given by  $\|G\|_{2,2} = \|\hat{G}\|_\infty$ . Each table entry also includes as a reminder the definition of each norm.

**Remark 3.54** Remember that the definitions of each norm simplify considerably in the SISO case! If you're feeling a bit overwhelmed by the table above, try rewriting it for the SISO case to gain some intuition for what's going on.

**Proof** As with the previous table, I will add a detailed proof for this result a little later in the quarter. Once again, in the meantime, you can refer to [43] for a proof of the full continuous-time MIMO case, and [12] for the continuous-time SISO case. You can find proofs of the discrete-time case in [10].  $\square$

### 3.3.3.3 State Space Methods for Computing System Norms

Theorem 3.15 provides us with a way of computing the induced  $p$ - $q$  system norms for  $p, q \in \{2, \infty\}$ . We note that two of the entries of each table,  $\|G\|_{2,2}$  and  $\|G\|_{2,\infty}$ , both still the rather complex formulas for the  $\mathcal{H}_\infty$  and  $\mathcal{H}_2$ -norms.

In general, such formulas *aren't* practical for control design. If, for instance, we design a controller and want to verify that it produces a system with a desirable  $\mathcal{H}_\infty$  or  $\mathcal{H}_2$  norm property, it's important that it's easy for us to compute the  $\mathcal{H}_\infty$  and  $\mathcal{H}_2$  norms. Looking even further down the line, if we want to design a controller which *optimizes* for the best  $\mathcal{H}_2$  or  $\mathcal{H}_\infty$  norm, we'd encounter a lot of trouble optimizing directly over their complex formulas!

How do we get around these problems? A useful insight is that, by converting back to state space, we can either *exactly* compute a  $\mathcal{H}_p$  norm (for  $p = 2$ ) or find a non-trivial upper bound on its value (for  $p = \infty$ ). What's more, these state space methods are amenable to computer implementation. First, we focus on computing the  $\mathcal{H}_2$  norm. Below, we prove the amazing result that we can *exactly* compute the  $\mathcal{H}_2$  norm of the transfer function of a system representation  $(A, B, C, 0)$  using a continuous-time Lyapunov equation.

**Theorem 3.16 (CTLE Computation of  $\mathcal{H}_2$ -Norm)** *Consider a continuous-time, LTI system representation  $(A, B, C, 0)$  with Hurwitz  $A$  matrix and transfer function  $\hat{G} \in \mathcal{RH}_\infty$ . The  $\mathcal{H}_2$ -norm of the transfer function is computed,*

$$\|\hat{G}\|_2 = \sqrt{\text{tr}(B^\top P B)}, \quad (3.276)$$

where  $P \in \mathbb{S}^n, P \succeq 0$  is the solution of the continuous-time Lyapunov equation,

$$A^\top P + P A = -C^\top C. \quad (3.277)$$

**Proof** Since  $\hat{G} \in \mathcal{RH}_\infty$  and  $A$  is Hurwitz, we may apply Parseval's theorem to compute the square of the  $\mathcal{H}_2$ -norm of  $\hat{G}$ . We have,

$$\|\hat{G}\|_2^2 = \int_{\mathbb{R}} \text{tr}(\hat{G}^*(j\omega)\hat{G}(j\omega))d\omega = \int_{\mathbb{R}_{\geq 0}} \text{tr}(G^\top(t)G(t))dt, \quad (3.278)$$

where  $G(\cdot) : \mathbb{R} \rightarrow \mathbb{R}^{p \times m}$  is the impulse response map of the system. Using the formula for the impulse response map of  $(A, B, C, 0)$ , it follows that

$$\|\hat{G}\|_2^2 = \int_{\mathbb{R}_{\geq 0}} \text{tr}(B^\top e^{A^\top t} C^\top C e^{A t} B)dt = \text{tr} \left( \int_{\mathbb{R}_{\geq 0}} B^\top e^{A^\top t} C^\top C e^{A t} B dt \right), \quad (3.279)$$

where we pull the trace out of the integral as a consequence of its linearity. Now, we pull  $B^\top$  and  $B$  out of the integral to get,

$$\|\hat{G}\|_2^2 = \text{tr} \left( B^\top \int_{\mathbb{R}_{\geq 0}} e^{A^\top t} C^\top C e^{At} dt B \right). \quad (3.280)$$

Amazingly, we recognize the term inside the integral as the solution  $P$  of the continuous-time Lyapunov equation  $A^\top P + PA = -C^\top C$ . Since  $C^\top C \succeq 0$ , it follows that  $P \succeq 0$ . One therefore has,

$$\|\hat{G}\|_2^2 = \text{tr}(B^\top P B) \implies \|\hat{G}\|_2 = \sqrt{\text{tr}(B^\top P B)}, \quad (3.281)$$

where the square root is real since  $P \succeq 0$ .  $\square$

**Exercise 3.27** Does an analogous result hold for the discrete-time case? If it does, state and prove the result. If not, explain what goes wrong.

Let's briefly discuss the implications of this result. Above, we showed that we can compute the  $\mathcal{H}_2$ -norm of a system using the solution to the Lyapunov equation  $A^\top P + PA = -C^\top C$ . Importantly, this equation is *linear* in  $P$ . This means that it is particularly easy to implement on the computer and to integrate into convex optimization problems.

Can we find an analogue of this result for the  $\mathcal{H}_\infty$ -norm? Due to the more complex analytical structure of the  $\mathcal{H}_\infty$ -norm, the best computationally tractable result we can obtain is an *upper bound* on the  $\mathcal{H}_\infty$ -norm.

**Theorem 3.17 (Upper Bound Certificate on  $\mathcal{H}_\infty$ -Norm)** *Consider a continuous-time, LTI system representation  $(A, B, C, 0)$  with Hurwitz  $A$  matrix and transfer function  $\hat{G} \in \mathcal{RH}_\infty$ . Let  $\gamma > 0$ . For the matrix,*

$$H_\gamma = \begin{bmatrix} A & \frac{1}{\gamma} B B^\top \\ -\frac{1}{\gamma} C^\top C & -A^\top \end{bmatrix}, \quad (3.282)$$

$\|\hat{G}\|_\infty < \gamma$  if and only if  $H$  has no eigenvalues on the  $j\omega$  axis.

*Remark 3.55*  $H$  is called a *Hamiltonian matrix*. We'll briefly discuss these matrices later in the course, in our study of optimal control.

*Remark 3.56* This result is a key component of one of the most famous papers in classical robust control. It was originally proven in [13], a paper that is now colloquially referred to as DGKF, in reference to its authors' last names. An accessible proof can be found in [43].

*Remark 3.57* To quote [13] on generalizations of some consequences of this result, "other extensions that we are way to cool to write up include the discrete-time and time-varying cases. Generalizations to infinite dimensional state-space models are the domain of the Mighty Thor."

**Proof (Sketch)** We provide a quick sketch of the logic of the proof, following [10]. A more thorough treatment of this result is given in [43]. First, we note that, by definition of the  $\mathcal{H}_\infty$  norm,  $\|\hat{G}\|_\infty < \gamma$  if and only if  $I - \frac{1}{\gamma^2} G^*(j\omega)G(j\omega)$  is invertible for all  $\omega \in \mathbb{R}$ . One may show that this is equivalent to  $[I - \frac{1}{\gamma^2} G^*(-s)G^*(s)]^{-1}$  having no poles on the imaginary axis. From here, one may establish that there exists a state space representation of  $[I - \frac{1}{\gamma^2} G^*(-s)G^*(s)]^{-1}$  for which the  $A$  matrix is  $H_\gamma$ . From this result, one may conclude the theorem.  $\square$



This result yields a computationally tractable method of computing an upper bound on the  $\mathcal{H}_\infty$ -norm of a transfer function. To test if  $\|\hat{G}\|_\infty < \gamma$  for some  $\gamma > 0$ , one calculates the eigenvalues of  $H_\gamma$ . Based on the results, one adjusts the value of  $\gamma$  and repeats the process.

### 3.3.4 Further Reading

There are a number of classic books in robust control that are good resources for further reading. For a not too theoretical overview of robust control, we recommend [42] and [32]. More in-depth treatments of robust control are found in [43] and [14]. The section on norms on signals, as well as SISO norms on systems, is based on [12]. The treatment of MIMO norms is primarily based on [43] and [14]. For more information on the discrete-time case, [10] is a useful resource. The treatment of scalar transfer functions via real-rational functions is based on that of [22], while the material on multivariable poles and zeros is based on [43] and [10].

### 3.3.5 Problems

**Problem 3.18 (The Real-Rational Equivalence Relation)** In this problem, we formalize the language of two rational functions being the same “up to multiplication by a nonzero  $p/p$ ” through the language of *equivalence relations*. An equivalence relation  $\mathcal{R}$  on a set  $S$  is a subset of  $S \times S$ , satisfying the following three axioms:

1. Identity:  $(x, x) \in \mathcal{R}$  for all  $x \in S$ .
2. Symmetry:  $(x, y) \in \mathcal{R} \Rightarrow (y, x) \in \mathcal{R}$ .
3. Transitivity:  $(x, y) \in \mathcal{R}$  and  $(y, z) \in \mathcal{R}$  implies  $(x, z) \in \mathcal{R}$ .

If two elements  $x, y$  satisfy  $(x, y) \in \mathcal{R}$ , they are said to be  $\mathcal{R}$ -related. For short, we denote this by  $x \sim y$ .

1. Consider the set  $S$  of scalar, real-rational functions. Show that the relation,

$$f_1(s) \sim f_2(s) \iff f_1 \text{ and } f_2 \text{ have the same coprime form,} \quad (3.283)$$

is an equivalence relation on  $S$ .

2. The equivalence class of an element  $x \in S$  is the set of all elements related to  $x$ ,

$$[x] = \{y \in S : x \sim y\}. \quad (3.284)$$

Using the equivalence relation defined in part (1), show that two real rational transfer functions  $\hat{G}_1, \hat{G}_2$  are “the same” in the sense of the section above if and only if  $[\hat{G}_1] = [\hat{G}_2]$ . This formalizes the notion of “being the same” we discussed in the section above.

3. Show that the poles and zeros of a given real-rational, scalar transfer function,  $\hat{G}$ , can be defined using only knowledge of the equivalence class,  $[\hat{G}]$ . Conclude that poles and zeros are “well-defined” concepts under the equivalence relation—i.e. that they don’t depend on the choice of element in the equivalence class.

**Problem 3.19 (Finite Impulse Response Systems)** In this problem, we'll consider a special type of discrete-time system, called a *finite impulse response (FIR)* system. A SISO discrete-time, LTI system with impulse response map  $H[\cdot] : \mathbb{Z} \rightarrow \mathbb{R}$  is said to be finite impulse response of horizon  $N$  if for all  $k > N$ ,  $H[k] = 0$ . That is,

$$H[0] = h_0, H[1] = h_1, \dots, H[N] = h_N, H[N+k] = 0, \forall k \geq 1, \quad (3.285)$$

where  $h_0, \dots, h_N \in \mathbb{R}$  are real constants.

1. Prove that the transfer function  $\hat{H}$  of a SISO, FIR system, must have all of its poles at the origin. Calculate an upper bound on the degree of the denominator of  $\hat{H}$  in terms of the FIR horizon.
2. Does a (SISO) transfer function  $\hat{G} \in \mathcal{RH}_\infty$  with all poles at the origin correspond to an FIR system? Provide a proof or counterexample to support your claim.
3. Let's consider input signals of length  $M \in \mathbb{Z}_{\geq 0}$ —input signals which are zero for all integers  $k \notin [0, M]$ . Calculate—as completely as you can—the 2-norm to 2-norm gain of a SISO FIR system with horizon  $N$ , over the set of length  $M$  inputs,

$$\sup_{\|u\|_{\ell_2}=1, \text{len}(u)=M} \|H * u\|_{\ell_2}. \quad (3.286)$$

Prove there exists a length  $M$  input  $u$  achieving the supremum. *You shouldn't have to use the table of system norms to answer this question!*

4. For the same system as above, calculate the infimum,

$$\inf_{\|u\|_{\ell_2}=1, \text{len}(u)=M} \|H * u\|_{\ell_2}. \quad (3.287)$$

Prove there exists a length  $M$  input  $u$  achieving the infimum. When (if ever) is this infimum zero?

**Problem 3.20 (Frequency Response [2])** In this problem, we'll study the frequency response of a SISO transfer function.

1. Consider a continuous-time, SISO system representation  $(A, B, C, D)$  in which  $A$  is Hurwitz. Show that, for an input signal  $u(t) = \sin(\omega_0 t + \phi)$ , the zero-state response of the system converges pointwise to the function,

$$y_{ss}(t) = |\hat{G}(j\omega_0)| \sin(\omega_0 t + \phi + \arg \hat{G}(j\omega_0)), \quad (3.288)$$

where  $\hat{G}$  is the transfer function of the system.

2. Extend your answer from part (a) to the MIMO case.

**Problem 3.21 (A  $\mathcal{H}_\infty$  Norm Identity [42])** Let  $\hat{G}$  be a continuous-time transfer function in  $\mathcal{RH}_\infty$ . Prove that,

$$\left\| \begin{bmatrix} \hat{G} \\ I \end{bmatrix} \right\|_\infty^2 = \|\hat{G}\|_\infty^2 + 1. \quad (3.289)$$

**Problem 3.22 (Scalar Parseval Theorem [42])** In this problem, we'll prove the scalar case of Parseval's theorem. Consider a function  $f \in L^2(\mathbb{R}_{\geq 0}, \mathbb{R})$ . Prove that,

$$\int_0^\infty (f(t))^2 dt = \frac{1}{2\pi} \int_{\mathbb{R}} |\hat{F}(j\omega)|^2 d\omega, \quad (3.290)$$

where  $\hat{F}(j\omega) = \int_0^\infty e^{-j\omega t} f(t) dt$ . *Hint: You may use without proof that*

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^\infty \hat{F}(j\omega) e^{j\omega t} d\omega. \quad (3.291)$$

*This is derived from the inverse Fourier transform of  $\hat{F}$ .*

**Problem 3.23 (Some System Norm Computations [12])**

1. Consider a SISO, LTI system with input  $u(\cdot)$ , output  $y(\cdot)$ , and transfer function,

$$\hat{G}(s) = \frac{s+2}{4s+1}, \quad (3.292)$$

Compute the norm  $\sup_{\|u\|_\infty=1} \|y\|_\infty$  and find an input signal  $u(\cdot)$  achieving this norm.

2. Compute the 1-norm of the impulse response map corresponding to the transfer function,

$$\hat{G}(s) = \frac{1}{\tau s + 1}, \quad \tau > 0. \quad (3.293)$$

**Problem 3.24 (Delay-Invariant Norms [12])** Recall that the transfer function for a time delay of time  $\tau$  is  $\hat{D}(s) = e^{-s\tau}$ . A norm  $\|\cdot\|$  on the space of transfer functions is delay-invariant if, for every transfer function  $\hat{G}$  with  $\|\hat{G}\| < \infty$  and every  $\tau > 0$ ,  $\|\hat{D}\hat{G}\| = \|\hat{G}\|$ . Is the 2-norm delay invariant? What about the  $\infty$ -norm? You may provide answers in the scalar (SISO) case.

**Problem 3.25 (Evaluating Potential Norms [12])** Consider the set  $C^1(\mathbb{R}, \mathbb{R})$  of continuously differentiable, scalar signals. Which of the following qualifies as a norm on  $C^1(\mathbb{R}, \mathbb{R})$ ?

$$(1) \sup_{t \in \mathbb{R}} |\dot{u}(t)| \quad (3.294)$$

$$(2) |u(0)| + \sup_{t \in \mathbb{R}} |\dot{u}(t)| \quad (3.295)$$

$$(3) \max\left\{\sup_{t \in \mathbb{R}} |u(t)|, \sup_{t \in \mathbb{R}} |\dot{u}(t)|\right\} \quad (3.296)$$

$$(4) \sup_{t \in \mathbb{R}} |u(t)| + \sup_{t \in \mathbb{R}} |\dot{u}(t)| \quad (3.297)$$



## Chapter 4

# Fundamental Limits on Linear Systems

Thus far in the course, we've focused primarily on *systems analysis*. We've studied the structure of solutions to linear systems and their representations, as well as their internal and external stability. With these tools at our disposal, we're ready to discuss where *control* comes into the picture.

An important first step in understanding how to design a controller for a linear system is to understand the *fundamental limitations* on the system. In this chapter, we'll focus on two fundamental limitations that are essential to the control design process: controllability and observability.

We'll begin with *controllability*—which considers whether we're able to control the system between given states and times—and then transition to observability—which considers if we're able to reconstruct the full state vector of the system from a set of measurements. Once we've established sharp conditions for controllability and observability for linear systems, we'll be ready to develop practical control design techniques and understand when and why they work. Let's begin!

### 4.1 Controllability

Imagine you're designing a control system to meet some specified performance objectives. Perhaps you'd like to take your system from state  $A$  at time  $t_0$  to state  $B$  at time  $t_1$  while minimizing the amount of energy exerted. Before you go about solving this control problem, it's important to ask—is the design task actually *feasible*?

Let's consider a couple of simple examples where the feasibility of the control design task becomes problematic. Suppose we'd like to design a control law  $u = k(x)$  to render the origin of the system,

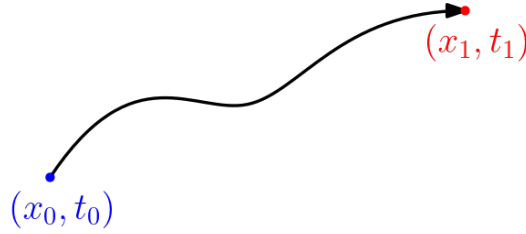
$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u, \quad (4.1)$$

globally exponentially stable. Is this task feasible? Based on our intuition, the answer seems to be no! The input  $u$  only appears on the  $x_1$  term, and the  $x_2$  term is *entirely* decoupled from  $x_1$ . Thus, at best, we should only be able to expect to stabilize the  $x_1$  component of

the system. No matter how clever we are with our control design, we have no control over the behavior of the  $x_2$  term!

This is a basic example of a *fundamental limit* on a control system. In this case, no matter what controller we pick, the structure of the system is such that we can't complete our desired control task. So, the question is—for which class of systems *can* we complete a meaningful range of control design tasks?

In order to precisely answer this question, we first need to come up with an idea of what a “meaningful” range of control design tasks actually refers to. A wide set of control design problems fall into the following broad category: we'll be given an initial state of the system  $x_0 \in \mathbb{R}^n$ , at a time  $t_0$ , and will be tasked with finding input  $u(\cdot) \in \mathcal{U}$  that takes the system to the state  $x_1 \in \mathbb{R}^n$  at a time  $t_1 > t_0$ .



**Fig. 4.1** Suppose we're given a pair of an initial state and time,  $(x_0, t_0)$ , and a pair of a final state and time,  $(x_1, t_1)$ . Can we find an input signal  $u(\cdot)$  that takes us from  $(x_0, t_0)$  to  $(x_1, t_1)$ ?

The question of whether or not we can construct an input  $u(\cdot)$  taking us between any two pairs  $(x_0, t_0)$  and  $(x_1, t_1)$  is rooted in the concept of *controllability*. As the course proceeds, we'll find that the idea of controllability is linked to a number of other important control design problems.

Let's sketch an outline for our study of controllability. First, we'll take the “transport between two states and times” problem which we sketched out above, and assign precise, mathematical language to each of its key components. Using this language, we'll formulate a formal definition of controllability. Following this, we'll specialize to the linear, time-invariant case, and come up with sharp *algebraic* characterizations of controllability. We'll then study how to decompose the state space of a system into controllable and uncontrollable components, and will consider an important system decomposition that arises from this decomposition. Finally, we'll finish up by studying some basic connections between controllability and optimal control.

#### 4.1.1 Defining Controllability & Reachability

Let's get started on the first component of our controllability program. We'll begin by assigning precise language to the simple controllability setup we outlined above: taking a system from a pair of an initial state and time,  $(x_0, t_0)$ , to a pair of a final state and time,  $(x_1, t_1)$ .

**Definition 4.1 (Event/Reachable Event)** Consider an I/O dynamical system  $\mathcal{D} = (\mathcal{U}, \mathcal{Y}, \Sigma, \varphi, r)$  on a time set  $\mathcal{T}$ .

1. **Event**: an event is a pair of a state and a time,  $(x, t) \in \Sigma \times \mathcal{T}$ .
2. **Reachable event**: consider the events  $(x_0, t_0)$  and  $(x_1, t_1)$ , where  $t_0 \leq t_1$ . One says that  $(x_1, t_1)$  *can be reached* from  $(x_0, t_0)$  if there exists an input signal  $u(\cdot) \in \mathcal{U}$  for which,

$$\varphi(t_1, t_0, x_0, u(\cdot)) = x_1. \quad (4.2)$$

If  $(x_1, t_1)$  can be reached from  $(x_0, t_0)$ , one writes  $(x_0, t_0) \rightsquigarrow (x_1, t_1)$ .

*Remark 4.1* Whenever we write  $(x_0, t_0) \rightsquigarrow (x_1, t_1)$ , we mean that there exists an input signal  $u(\cdot) \in \mathcal{U}$  for which  $\varphi(t_1, t_0, x_0, u(\cdot)) = x_1$ . Note that we *do not* claim that this input signal is unique! All that we ask is that *some* signal exists.

Using the notion of a reachable event, we state a precise definition for controllability. The basic idea of this definition is the following: if, for given times  $t_0 \leq t_1$ , we can reach *any* state  $x_1$  at  $t_1$  from *any* state  $x_0$  at  $t_0$ , we declare the system to be controllable on  $[t_0, t_1]$ .

**Definition 4.2 (Controllability)** Consider an I/O dynamical system  $\mathcal{D}$  with state space  $\Sigma$  and time set  $\mathcal{T}$ . Let  $t_0, t_1 \in \mathcal{T}$  with  $t_0 \leq t_1$ . The system  $\mathcal{D}$  is said to be *controllable* on the interval  $[t_0, t_1] \subseteq \mathcal{T}$  if, for all  $x_0, x_1 \in \Sigma$ ,  $(x_0, t_0) \rightsquigarrow (x_1, t_1)$ .

*Remark 4.2* Here,  $[t_0, t_1] \subseteq \mathcal{T}$  refers to the interval  $[t_0, t_1] \cap \mathcal{T}$ . It's important to note that  $[t_0, t_1]$  is *not* necessarily an interval in  $\mathbb{R}$  when used in this context! Rather, it is an interval in the time set  $\mathcal{T}$ .

*Remark 4.3* It's important to note—the definition of controllability necessitates that we be at  $x_1$  *exactly* at time  $t_1$ ! If we reach  $x_1$  before time  $t_1$ , and don't remain there, we do not satisfy the conditions for controllability.

*Remark 4.4* Controllability is something that depends only on the *state* behavior of the system! As can be seen in the definition, whether or not a system is controllable is captured entirely by the structure of the state transition map. As such, the readout and I/O maps  $r$  and  $\rho$  are not of interest in the context of controllability.

This definition formalizes the basic concept of controllability that we sketched out above. Namely, if we can go from any state at time  $t_0$  to any state at time  $t_1$ , the system is said to be controllable on the interval  $[t_0, t_1]$ .

How can characterize the controllability of an I/O dynamical system? As you might recall, it's generally challenging to say interesting things about *abstract* dynamical systems—the unstructured nature of these systems means that analyzing even the simplest concepts can be a fruitless pursuit. In order to get a feel for the structure of controllability, we therefore impose further structure on the system  $\mathcal{D}$ . Since the notion of controllability depends almost entirely on the structure of the state transition map, a good place to start is in imposing regularity conditions on the state transition map.

Recall that, in Chapter 2, we declared an input/output system  $\mathcal{D} = (\mathcal{U}, \mathcal{Y}, \Sigma, \varphi, r)$  to be *linear* if each of the spaces  $\mathcal{U}, \mathcal{Y}, \Sigma$  was a vector space, and if the I/O map  $\rho$  was linear in  $\Sigma \times \mathcal{U}$ . Likewise, we declared the system to be *time-invariant* if the spaces  $\mathcal{U}, \mathcal{Y}$  and the I/O map were *delay-invariant*. Since, in the case of controllability, we're interested in the behavior of the *state* and not the behavior of the *output*, we specialize the definitions of linearity and time-invariance to the state transition map.

**Definition 4.3 (Linear/Time-Invariant State Transition Map)** Consider an I/O dynamical system  $\mathcal{D} = (\mathcal{U}, \mathcal{Y}, \Sigma, \varphi, r)$ . The state transition map  $\varphi$  is:

1. **Linear:** if each of  $\mathcal{U}, \Sigma$  are vector spaces over a common field  $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$  and  $\varphi$  is linear in  $\Sigma \times \mathcal{U}$ . That is, for all  $\alpha, \beta \in \mathbb{K}$ ,  $x_0, \hat{x}_0 \in \Sigma$ , and  $u(\cdot), \hat{u}(\cdot) \in \mathcal{U}$ , and  $t_0 \leq t_1 \in \mathcal{T}$ ,

$$\varphi(t_1, t_0, \alpha x_0 + \beta \hat{x}_0, \alpha u(\cdot) + \beta \hat{u}(\cdot)) = \alpha \varphi(t_1, t_0, x_0, u_0) + \beta \varphi(t_1, t_0, \hat{x}_0, \hat{u}_0). \quad (4.3)$$

2. **Time-invariant:** if the input-signal space  $\mathcal{U}$  is a delay-invariant set and the state transition map  $\varphi$  is delay-invariant. That is, for all  $t_0, t_1, \tau \in \mathcal{T}$  with  $t_0 \leq t_1$ , all  $x_0 \in \Sigma$ , and  $u(\cdot) \in \mathcal{U}$ , the state transition map satisfies,

$$\varphi(t_1, t_0, x_0, u(\cdot)) = \varphi(t_1 + \tau, t_0 + \tau, x_0, T_\tau(u(\cdot))), \quad (4.4)$$

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

If  $\varphi$  is both linear and time-invariant, it is said to be linear, time-invariant (LTI).

*Remark 4.5* If  $\varphi$  is linear on a set of vector spaces over a field  $\mathbb{K}$ , one says that “ $\varphi$  is linear over  $\mathbb{K}$ .”

*Remark 4.6* Take a look back at the definitions of linearity and time invariance from Chapter 2. You'll notice that, in the above, *all that we're doing* is replacing the I/O map  $\rho$  with the state transition map  $\varphi$ . Under this observation, we note that we can alternatively express that the state transition map is linear and time-invariant by requesting that the I/O system with readout map  $r_\Sigma(t, x, u) = x$  is linear and time-invariant in the input/output sense. This is a nice fact to verify to check your understanding!

What are some basic examples of systems with linear and time-invariant state transition maps? Based on the structure of their solutions, every linear, time-varying state space representation has a linear state transition map, and every linear, time-invariant state space representation has a linear, time-invariant state transition map.

**Exercise 4.1** Verify that a continuous or discrete-time linear, time-varying state space representation  $(A(\cdot), B(\cdot), C(\cdot), D(\cdot))$  has a linear state transition map. Verify that a continuous or discrete-time linear, time-invariant state space representation  $(A, B, C, D)$  has a linear, time-invariant state transition map.

Before we jump into an analysis of controllability for linear and time-invariant systems, we introduce one more definition. In the case where the state space  $\Sigma$  of the system is a *vector space* (as is the case for a system with a *linear* state transition map), we can define a more specialized form of controllability called *reachability*.

**Definition 4.4 (Reachability)** Consider an I/O dynamical system  $\mathcal{D}$  on a time set  $\mathcal{T}$ . Suppose the state space  $\Sigma$  of  $\mathcal{D}$  is a vector space. For times  $t_0, t_1 \in \mathcal{T}$  with  $t_0 \leq t_1$ , the system is said to be *reachable* on  $[t_0, t_1]$  if for all  $x \in \Sigma$ ,  $(0, t_0) \rightsquigarrow (x, t_1)$ .

*Remark 4.7* It's important to note that we *cannot* in general define reachability for a system on an arbitrary state space  $\Sigma$ , as an arbitrary set  $\Sigma$  is not guaranteed to have a zero element. Since a vector space always contains a zero vector, the definition of reachability is well-posed when  $\Sigma$  is a vector space.

Thus, a system on a vector space is said to be *reachable* on  $[t_0, t_1]$  if we can reach any state at time  $t_1$  from the origin at time  $t_0$ . Although it might initially seem like reachability is a much weaker property than controllability, we find that, for linear systems, the two are in fact *equivalent*. We state this and two other facts in the following proposition.



**Proposition 4.1 (Controllability and Reachability for LTI Systems)** *Consider an I/O dynamical system  $\mathcal{D} = (\mathcal{U}, \mathcal{Y}, \Sigma, \varphi, r)$  on a time set  $\mathcal{T}$ .*

1. Linearity: *if  $\varphi$  is linear over a field  $\mathbb{K}$  and  $(x_0, t_0) \rightsquigarrow (x_1, t_1)$ ,  $(\hat{x}_0, t_0) \rightsquigarrow (\hat{x}_1, t_1)$ , then,*

$$(\alpha x_0 + \beta \hat{x}_0, t_0) \rightsquigarrow (\alpha x_1 + \beta \hat{x}_1, t_1), \quad \forall \alpha, \beta \in \mathbb{K}. \quad (4.5)$$

2. Time-invariance: *if  $\varphi$  is time-invariant, then for all  $\tau \in \mathcal{T}$ ,*

$$(x_0, t_0) \rightsquigarrow (x_1, t_1) \iff (x_0, t_0 + \tau) \rightsquigarrow (x_1, t_1 + \tau). \quad (4.6)$$

3. Controllability iff reachability: *if  $\varphi$  is linear, then for any pair of times  $t_0 \leq t_1 \in \mathcal{T}$ ,  $\mathcal{D}$  is reachable on  $[t_0, t_1]$  if and only if it is controllable on  $[t_0, t_1]$ .*

**Proof** First, we show item (1). Consider  $x_0, \hat{x}_0, x_1, \hat{x}_1 \in \Sigma$  and  $t_0 \leq t_1 \in \mathcal{T}$ . Fix a pair of scalars  $\alpha, \beta \in \mathbb{K}$ . We want to show that,

$$(x_0, t_0) \rightsquigarrow (x_1, t_1), (\hat{x}_0, t_0) \rightsquigarrow (\hat{x}_1, t_1) \implies (\alpha x_0 + \beta \hat{x}_0, t_0) \rightsquigarrow (\alpha x_1 + \beta \hat{x}_1, t_1). \quad (4.7)$$

If  $(x_0, t_0) \rightsquigarrow (x_1, t_1)$  and  $(\hat{x}_0, t_0) \rightsquigarrow (\hat{x}_1, t_1)$ , there exist input signals  $u(\cdot), \hat{u}(\cdot) \in \mathcal{U}$  for which

$$\varphi(t_1, t_0, x_0, u(\cdot)) = x_1, \quad \varphi(t_1, t_0, \hat{x}_0, \hat{u}(\cdot)) = \hat{x}_1. \quad (4.8)$$

By linearity of the state transition map, it follows that,

$$\varphi(t_1, t_0, \alpha x_0 + \beta \hat{x}_0, \alpha u(\cdot) + \beta \hat{u}(\cdot)) = \alpha \varphi(t_1, t_0, x_0, u(\cdot)) + \beta \varphi(t_1, t_0, \hat{x}_0, \hat{u}(\cdot)) \quad (4.9)$$

$$= \alpha x_1 + \beta \hat{x}_1. \quad (4.10)$$

Since there exists an input delivering  $\alpha x_0 + \beta \hat{x}_0$  at time  $t_0$  to  $\alpha x_1 + \beta \hat{x}_1$  at time  $t_1$ , we conclude that  $(\alpha x_0 + \beta \hat{x}_0, t_0) \rightsquigarrow (\alpha x_1 + \beta \hat{x}_1, t_1)$ . This completes the proof of item (1).

Now, we show item (2). Fix a time  $\tau \in \mathcal{T}$ . Suppose one has  $(x_0, t_0) \rightsquigarrow (x_1, t_1)$ . Then, there exists an input  $u(\cdot)$  for which  $\varphi(t_1, t_0, x_0, u(\cdot)) = x_1$ . Applying the definition of time-invariance, we have that,

$$\varphi(t_1 + \tau, t_0 + \tau, x_0, T_\tau(u)) = x_1, \quad (4.11)$$

where  $T_\tau$  is the delay map,  $T_\tau(u)(t) = u(t - \tau)$ . Thus, the input signal  $T_\tau(u)$  delivers the system from state  $x_0$  at time  $t_0 + \tau$  to  $x_1$  at time  $t_1 + \tau$ . Thus,  $(x_0, t_0) \rightsquigarrow (x_1, t_1)$  implies  $(x_0, t_0 + \tau) \rightsquigarrow (x_1, t_1 + \tau)$ . For the other direction, if  $(x_0, t_0 + \tau) \rightsquigarrow (x_1, t_1 + \tau)$  for all  $\tau \in \mathcal{T}$ , taking  $\tau = 0$  leads us to conclude  $(x_0, t_0) \rightsquigarrow (x_1, t_1)$  (recall that, in Chapter 2, we required that the time set  $\mathcal{T}$  contain zero—taking  $\tau = 0$  is therefore allowed). This completes the proof of item (2).

Finally, we show item (3). Suppose the system is controllable on  $[t_0, t_1]$ . Then, for all  $x_0, x_1 \in \Sigma$ ,  $(x_0, t_0) \rightsquigarrow (x_1, t_1)$ . Taking  $x_0 = 0$ , we conclude that the system is reachable on  $[t_0, t_1]$ . Now, suppose the system is reachable on  $[t_0, t_1]$ . Fix a pair of states  $x_0, x_1 \in \Sigma$ . We wish to show that  $(x_0, t_0) \rightsquigarrow (x_1, t_1)$ . By reachability, we know

$$(0, t_0) \rightsquigarrow (x_1 - \varphi(t_1, t_0, x_0, 0), t_1). \quad (4.12)$$

This implies the existence of an input  $u(\cdot) \in \mathcal{U}$  for which,

$$\varphi(t_1, t_0, 0, u(\cdot)) = x_1 - \varphi(t_1, t_0, x_0, 0) \quad (4.13)$$

$$\varphi(t_1, t_0, 0, u(\cdot)) + \varphi(t_1, t_0, x_0, 0) = x_1 \quad (4.14)$$

$$\varphi(t_1, t_0, x_0, u(\cdot)) = x_1, \quad (4.15)$$

where we use linearity to combine the sum of the state transition maps. We conclude that  $(x_0, t_0) \rightsquigarrow (x_1, t_1)$ . The system is therefore controllable on  $[t_0, t_1]$ .  $\square$

Let's summarize the results of the proposition above. We showed that, for a system with a linear state transition map, any linear combination of reachable events is reachable. Then, we showed that, for a time-invariant system, an event reachable on  $[t_0, t_1]$  implies the event will be reachable on  $[t_0 + \tau, t_1 + \tau]$  for *any*  $\tau \in \mathcal{T}$ . Notably, this implies that, for a time-invariant system, analyzing the reachability of an event on  $[0, t_1 - t_0]$  is equivalent to analyzing reachability on  $[t_0, t_1]$ . Finally, we showed that, for a system with a linear state transition map, the system is controllable if and only if it is reachable. With the different pieces of this result in mind, we define a set called the *reachable subspace*.

**Definition 4.5 (Reachable Subspace)** Let  $\mathcal{D} = (\mathcal{U}, \mathcal{Y}, \Sigma, \varphi, r)$  be an I/O dynamical system with a linear state transition map  $\varphi$ . The *reachable subspace* of  $\mathcal{D}$  in time  $T \in \mathcal{T}_{\geq 0}$  is the set,

$$\mathcal{R}_T = \{x \in \Sigma : (0, 0) \rightsquigarrow (x, T)\}. \quad (4.16)$$

*Remark 4.8* The set  $\mathcal{R}_T$  simply contains all points we can reach at *exactly* time  $T$ . One can define the set of all points we can reach *up until* time  $T$  as  $\mathcal{R}_{\leq T} = \cup_{t \in [0, T]} \mathcal{R}_t$ . Later, we'll find that, for a linear, time-invariant system representation,  $\mathcal{R}_{\leq T} = \mathcal{R}_T$  for all  $T > 0$ .

*Remark 4.9* For a system with a linear, time-invariant state transition map,  $\mathcal{R}_T = \{x \in \Sigma : (0, 0) \rightsquigarrow (x, T)\}$  is equal to  $\{x \in \Sigma : (0, t_0) \rightsquigarrow (x, t_1)\}$  for any  $t_0, t_1$  satisfying  $t_1 - t_0 = T$ . As the linear, time-invariant case will be of primary interest to us in this section, we only consider reachable subspaces defined on intervals of the form  $[0, T]$ .

Using Proposition 4.1, we confirm that, when  $\varphi$  is a linear,  $\mathcal{R}_T$  is indeed a subspace of the state space  $\Sigma$ .

**Corollary 4.1 (Reachable Subspace is a Subspace)** Consider an I/O dynamical system  $\mathcal{D}$  with a linear state transition map  $\varphi$ . For any  $T \in \mathcal{T}_{\geq 0}$ , the reachable subspace  $\mathcal{R}_T$  of  $\mathcal{D}$  is a subspace of  $\Sigma$ .

**Proof** Suppose  $x_0, \hat{x}_0 \in \mathcal{R}_T$ . Then,  $(0, 0) \rightsquigarrow (x_0, T)$  and  $(0, 0) \rightsquigarrow (\hat{x}_0, T)$ . By Proposition 4.1, we conclude that for any scalars  $\alpha, \beta \in \mathbb{K}$ ,  $(0, 0) \rightsquigarrow (\alpha x_0 + \beta \hat{x}_0, T)$ . We conclude that  $\alpha x_0 + \beta \hat{x}_0 \in \mathcal{R}_T$ , for any  $\alpha, \beta \in \mathbb{K}$ , which implies that  $\mathcal{R}_T$  is a subspace of  $\Sigma$ .  $\square$

Using the language of reachability and reachable subspaces, we state three equivalent ways of stating that a system with a linear, time-invariant state transition map is controllable.

**Corollary 4.2 (Controllability, Stated Three Ways)** Consider a dynamical system  $\mathcal{D}$  with a linear, time-invariant state transition map  $\varphi$ . For each fixed pair of times  $t_0, t_1 \in \mathcal{T}$ ,  $t_0 \leq t_1$ , the following are equivalent:

1. Controllability: the system is controllable on  $[t_0, t_1]$ .
2. Reachability: the system is reachable on  $[0, t_1 - t_0]$ .
3. Reachable subspace: the reachable subspace satisfies  $\mathcal{R}_{t_1 - t_0} = \Sigma$ .

**Proof** We'll show  $(1) \Rightarrow (2) \Rightarrow (3) \Rightarrow (1)$  to prove that the three statements are equivalent. First, we know from Proposition 4.1 that controllability on  $[t_0, t_1]$  implies reachability on  $[t_0, t_1]$ . Since the system is time-invariant, reachability on  $[t_0, t_1]$  implies reachability on  $[t_0 + \tau, t_1 + \tau]$  for any  $\tau \in \mathcal{T}$ . Taking  $\tau = -t_1$ , it follows that the system is reachable on  $[0, t_1 - t_1]$ . Thus, we conclude that  $(1) \Rightarrow (2)$ .

Now, we show that  $(2) \Rightarrow (3)$ . If the system is reachable on  $[0, t_1 - t_0]$ , then for any  $x \in \Sigma$ ,  $(0, 0) \rightsquigarrow (x, t_1 - t_0)$ . Thus, by definition of the reachable subspace, we have  $\mathcal{R}_{t_1 - t_0} = x$ . It follows that  $(2) \Rightarrow (3)$ .

Finally, we show that  $(3) \Rightarrow (1)$ . Since  $\mathcal{R}_{t_1 - t_0} = \Sigma$  implies that the system is reachable on  $[0, t_1 - t_0]$ , we conclude that the system is controllable on  $[0, t_1 - t_0]$ . Since controllability on  $[0, t_1 - t_0]$  implies controllability  $[t_0, t_1]$  (as the system is time-invariant), we conclude that  $(3) \Rightarrow (1)$ . This completes the proof.  $\square$

Motivated by this result, in the remainder of this section, we'll concern ourselves with answering the three equivalent questions:

1. When is a system with an LTI state transition map controllable on  $[t_0, t_1]$ ,  $t_1 - t_0 = T$ ?
2. When is a system with an LTI state transition map reachable on  $[0, T]$ ?
3. When does a system with an LTI state transition map satisfy  $\mathcal{R}_T = \Sigma$ ?

In order to provide sharp, algebraic answers to each of these questions, we study the controllability and reachability of linear, time-invariant state space representations.

#### 4.1.2 Controllability of Linear, Time-Invariant Systems

In previous sections, after introducing a concept for an abstract dynamical system, we worked on identifying a sharp, algebraic characterization of the concept for a linear, time-invariant system representation. When studying stability, for instance, after introducing general definitions, we proved that we can get an exact characterization of global exponential stability through the eigenvalues of the  $A$  matrix of the system representation.

We now seek to perform a similar characterization of the controllability of linear, time-invariant state space representations. In particular, we'd like to identify a sharp characterization of controllability through a purely algebraic condition as easy to check as the eigenvalue condition for stability. As usual, we'll split this task into the continuous and discrete-time cases. We'll tackle the more challenging continuous-time case first, and will then proceed to study the analogous discrete-time theory.

##### 4.1.2.1 The Continuous-Time Case

We begin with the case of a continuous-time, LTI state space representation. Recall that such a system representation has a state equation,

$$\dot{x}(t) = Ax(t) + Bu(t), \quad (4.17)$$

where  $x(t) \in \mathbb{R}^n$  and  $u(t) \in \mathbb{R}^m$ . Given any pair of times  $t_0 < t_1 \in \mathbb{R}$ , we'd like to identify necessary and sufficient algebraic conditions under which the system is controllable on  $[t_0, t_1]$ .

Since controllability on  $[t_0, t_1]$  is equivalent to reachability on  $[0, t_1 - t_0]$  for a system with an LTI state transition map, we'll begin the search for algebraic conditions for controllability by studying the simpler problem of reachability on an interval  $[0, T]$ . Fix an arbitrary state  $x \in \mathbb{R}^n$  and an arbitrary time  $T > 0$ . For this state and time, let's try solving for an input signal  $u(\cdot)$  satisfying  $\varphi(T, 0, 0, u(\cdot)) = x$ . Since the existence of such a signal is required for the system to be reachable, this study will provide us with some insight into the reachability problem. Based on the formula for the continuous-time, LTI state transition map, we'd like to identify an input signal  $u(\cdot)$  satisfying,

$$x = \varphi(T, 0, 0, u(\cdot)) = \int_0^T e^{A(T-\tau)} B u(\tau) d\tau. \quad (4.18)$$

How can we solve this equation for an input signal  $u(\cdot) \in \mathcal{U}$ ? Let's see if we can simplify this expression to a form in which we can identify  $u(\cdot)$  via a simple matrix inversion. In this spirit, we consider the input,

$$u(t) = B^\top \exp(A^\top (T - t)) M v, \quad (4.19)$$

which takes the transpose of the integrand term  $e^{A(T-\tau)} B$  and multiplies it by a matrix  $M$  and a vector  $v$  which we have yet to determine. Plugging this in, we find,

$$\int_0^T e^{A(T-\tau)} B u(\tau) d\tau = \int_0^T \left[ e^{A(T-\tau)} B B^\top e^{A^\top (T-\tau)} \right] M v d\tau \quad (4.20)$$

$$= \left[ \int_0^T e^{A(T-\tau)} B B^\top e^{A^\top (T-\tau)} d\tau \right] M v. \quad (4.21)$$

Now, we have a square matrix, contained in an integral, multiplied by  $Mv$ . If we assume that the entire integral term is *invertible*, then we can pick  $M$  as the inverse of the integral term and  $v$  as the desired final state,  $x$ . This choice yields,

$$\left[ \int_0^T e^{A(T-\tau)} B B^\top e^{A^\top (T-\tau)} d\tau \right] M v \quad (4.22)$$

$$= \left[ \int_0^T e^{A(T-\tau)} B B^\top e^{A^\top (T-\tau)} d\tau \right] \left[ \int_0^T e^{A(T-\tau)} B B^\top e^{A^\top (T-\tau)} d\tau \right]^{-1} x = x. \quad (4.23)$$

Thus, provided we're able to *invert* the integral term, we'll be able to solve the reachability (and therefore controllability) problem. This observation leads to the following definition.

**Definition 4.6 (Controllability Gramian)** Consider a continuous-time, LTI system representation  $(A, B, C, D)$ . The controllability Gramian is the map  $W_c : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^{n \times n}$ , defined

$$W_c(T) := \int_0^T e^{A\tau} B B^\top e^{A^\top \tau} d\tau. \quad (4.24)$$

*Remark 4.10* Notice that, in this definition, we've dropped the  $T - \tau$  from our derivation above in favor of  $\tau$ —this is simply a change of variables from the integral we derived above.

*Remark 4.11* Controllability Gramians are members of a wider class of matrices called *Gram matrices*. As the set of Gram matrices is named after the mathematician Jørgen Pedersen Gram, one typically capitalizes the terms *Gram* and *Gramian*.

**Exercise 4.2** Confirm that, for any LTI system representation, the controllability Gramian  $W_c(T)$  is symmetric and positive semidefinite ( $x^\top W_c(T)x \geq 0$  for all  $x \in \mathbb{R}^n$ ) for all  $T > 0$ .

Above, we sketched out a basic proof of the fact that, if the controllability Gramian is invertible (in fact positive definite) at time  $T$ , then the system is reachable on  $[0, T]$ . Now, we'll state and prove a much stronger result. We'll show that positive definiteness of the controllability Gramian is both necessary *and* sufficient for reachability. Further, we'll prove that if the Gramian is positive definite at one time  $T_0 > 0$ , then it must be positive definite for *all*  $T > 0$ .

**Theorem 4.1 (Gramian Characterization of Controllability)** *Consider a continuous-time, LTI system representation with state equation  $\dot{x}(t) = Ax(t) + Bu(t)$ . The following statements are equivalent:*

1. *There exists a time  $T > 0$  at which the controllability gramian  $W_c(T)$  is positive definite.*
2. *For any  $T > 0$ , the controllability gramian  $W_c(T)$  is positive definite.*
3. *For any  $T > 0$ , the system is reachable on  $[0, T]$ .*
4. *For any  $[t_0, t_1] \subseteq \mathbb{R}$ ,  $t_0 < t_1$ , the system is controllable on  $[t_0, t_1]$ .*
5. *There exists an interval  $[t_0, t_1] \subseteq \mathbb{R}$ ,  $t_0 < t_1$ , on which the system is controllable.*

**Proof** To prove these statements are equivalent, we'll show,

$$(1) \Rightarrow (2) \Rightarrow (3) \Rightarrow (4) \Rightarrow (5) \Rightarrow (1). \quad (4.25)$$

By following this “circle of implications,” we can conclude that each of (1) through (5) is in fact equivalent. First, we show that (1)  $\Rightarrow$  (2). Suppose there exists a time  $T > 0$  at which the controllability Gramian  $W_c(T)$  is positive definite. We'd like to show that the Gramian is positive definite for *all*  $T > 0$ . For contradiction, assume there exists a time  $T' > 0$  at which the Gramian is *not* positive definite. Since the Gramian is always at least positive semidefinite, this implies  $\text{rank}(W_c(T')) < n$ . Thus, there exists a nonzero vector  $v \in \mathbb{R}^n$  for which,

$$0 = v^\top W_c(T')v \quad (4.26)$$

$$= v^\top \int_0^{T'} e^{A\tau} B B^\top e^{A^\top \tau} d\tau v \quad (4.27)$$

$$= \int_0^{T'} (v^\top e^{A\tau} B)(B^\top e^{A^\top \tau} v) d\tau \quad (4.28)$$

$$= \int_0^{T'} \|B^\top e^{A^\top \tau} v\|_2^2 d\tau. \quad (4.29)$$

Since the 2-norm is nonnegative, in order for this integral to equal zero, we must have,

$$B^\top e^{A^\top t} v = 0, \quad \forall t \in [0, T']. \quad (4.30)$$

This implies that, for all  $t \in (0, T')$ ,  $\frac{d^i}{dt^i} B^\top e^{A^\top \tau} v = 0$  for all  $i \in \mathbb{Z}_{\geq 0}$ <sup>1</sup>. Thus,

$$B^\top e^{A^\top t} (A^\top)^i v = 0, \quad \forall t \in (0, T'), \quad i \in \mathbb{Z}_{\geq 0}. \quad (4.31)$$

Taking the limit of this equality as  $t \rightarrow 0^+$ , we conclude that  $B^\top (A^\top)^i v = 0$  for all  $i \in \mathbb{Z}_{\geq 0}$ . From the series definition of the matrix exponential, we then conclude that  $B^\top e^{A^\top t} v = 0$  for all  $t \in \mathbb{R}$ —not just for  $t \in [0, T']$ . But, this implies that  $0 = v^\top W_c(T) v$ . Since we assumed that  $W_c(T) \succ 0$ , we've reached a contradiction! We conclude that  $W_c(T)$  must be positive definite for all  $T > 0$ .

Now, we show (2)  $\Rightarrow$  (3). We already sketched out the argument for this implication above, but will repeat it here for the sake of completeness. Suppose the controllability Gramian  $W_c(T)$  is positive definite for all  $T > 0$ . We will show that, for any  $T > 0$ , the system is *reachable* on  $[0, T]$ . Fix a time  $T > 0$  and a state  $x \in \mathbb{R}^n$ . To show reachability, we must show there exists an input  $u(\cdot) \in \mathcal{U}$  for which  $\varphi(T, 0, 0, u(\cdot)) = x$ . Consider the input signal,  $u(\cdot) : [0, T] \rightarrow \mathbb{R}^m$ , defined,

$$u(t) = B^\top e^{A^\top (T-t)} W_c(T)^{-1} x. \quad (4.32)$$

Since  $W_c(T)$  is positive definite for all  $T > 0$ , this signal is well-defined. Note that, to ensure that this signal formally belongs to  $\mathcal{U} = PC(\mathbb{R}, \mathbb{R}^m)$ , we harmlessly identify it with a signal equal to  $u(t)$  on  $[0, T]$  and zero outside  $[0, T]$ . This extends the domain of the signal to all of  $\mathbb{R}$ , and makes it a member of  $\mathcal{U}$ . For such an input signal, we have,

$$\varphi(T, 0, 0, u(\cdot)) = \int_0^T e^{A(T-\tau)} B u(\tau) d\tau \quad (4.33)$$

$$= \int_0^T \left[ e^{A(T-\tau)} B B^\top e^{A^\top (T-\tau)} \right] W_c(T)^{-1} x d\tau \quad (4.34)$$

$$= \left[ \int_0^T e^{A\tau} B B^\top e^{A^\top \tau} d\tau \right] W_c(T)^{-1} x \quad (4.35)$$

$$= W_c(T) W_c(T)^{-1} x = x, \quad (4.36)$$

where we perform the change of variables from  $\tau$  to  $T - \tau$  to simplify the integral. We conclude that the proposed input signal delivers the system from  $x_0 = 0$  at time 0 to  $x$  at time  $T$ . Since the state  $x$  and the time  $T$  were chosen arbitrarily, we conclude that (2)  $\Rightarrow$  (3).

In Proposition 4.1, we already showed that (3)  $\Rightarrow$  (4). Additionally, the implication (4)  $\Rightarrow$  (5) is trivial, as controllability on every interval automatically implies controllability on a single interval. As such, all that remains is to show that (5)  $\Rightarrow$  (1). Suppose there exists a pair of times  $t_0 < t_1 \in \mathbb{R}$  for which the system is controllable on  $[t_0, t_1]$ . By Proposition 4.1, this is equivalent to controllability on  $[0, T]$ , where  $T = t_1 - t_0$ . Let's show that  $W_c(T)$  is positive definite. For contradiction, suppose that  $W_c(T)$  is *not* positive definite. Then, there exists a nonzero vector  $v \in \mathbb{R}^n$  for which

$$0 = v^\top W_c(T) v = \int_0^T v^\top e^{A\tau} B B^\top e^{A^\top \tau} v d\tau = \int_0^T \|B^\top e^{A^\top \tau} v\|_2^2 d\tau. \quad (4.37)$$

<sup>1</sup> Later in this section, we'll prove that it's sufficient to only evaluate the first  $n - 1$  derivatives. This is a consequence of a result called the *Cayley-Hamilton theorem*, which we'll also prove below.

Since the norm is nonnegative, we conclude that  $v^\top e^{At}B = 0$  for all  $t \in [0, T]$ . Since the system is controllable on  $[0, T]$ , we must have that  $(x_0, 0) \rightsquigarrow (0, T)$ , for any  $x_0 \in \mathbb{R}^n$ . Thus, for each  $x_0 \in \mathbb{R}^n$ , there exists an input  $u(\cdot)$  for which,

$$0 = e^{AT}x_0 + \int_0^T e^{A(T-\tau)}Bu(\tau)d\tau. \quad (4.38)$$

Left-multiplying both sides of the equation above by  $v^\top$ , we find,

$$0 = v^\top e^{AT}x_0 + \int_0^T v^\top e^{A(T-\tau)}Bu(\tau)d\tau = v^\top e^{AT}x_0, \quad (4.39)$$

where the integral term vanishes since  $v^\top e^{At}B = 0$ . Picking an initial condition  $x_0 = e^{-AT}v$ , we get  $0 = v^\top v$ , which implies  $v$  is the zero vector. Since  $v \neq 0$ , this is a contradiction. We conclude that  $W_c(T)$  is positive definite at  $T > 0$ . As there exists a time  $T > 0$  at which  $W_c(T)$  is positive definite, we conclude that (5)  $\Rightarrow$  (1).  $\square$

This result has a number of *fundamentally important* consequences for the controllability of linear, time-invariant state space representations. First, we showed that we can *completely* characterize the controllability of a linear system using the *controllability Gramian*, a matrix which depends on the  $A$  and  $B$  matrices of the system representation. We also showed that, if the controllability Gramian is positive definite at *one time*, then it is positive definite for all time! Consequently, we were able to conclude that a linear, time-invariant system is controllable on *any* interval if and only if there exist a *single* interval on which it's controllable.

For a linear, time-invariant system, it therefore makes sense to talk about controllability as a property of the *system*, rather than a property of the system and a time interval. As a consequence of this, one often says that the *system* with state equation,

$$\dot{x}(t) = Ax(t) + Bu(t), \quad (4.40)$$

is controllable, without any reference to a time interval. Since the state equation  $\dot{x}(t) = Ax(t) + Bu(t)$  determines controllability on any interval and is itself uniquely determined by  $A$  and  $B$ , we make the following definition.

**Definition 4.7 (Controllable Pair)** A pair of matrices  $(A, B)$ , where  $A \in \mathbb{R}^{n \times n}$  and  $B \in \mathbb{R}^{n \times m}$ , is said to be a *controllable pair* if there exists an interval on which the system with state equation  $\dot{x}(t) = Ax(t) + Bu(t)$  is controllable.

**Exercise 4.3** Convince yourself that this definition is well-posed—that it's justified to not mention an interval and that there is no ambiguity in declaring controllability to be a property of the  $A$  and  $B$  matrices.

Let's focus in on a few of steps we used in the proof of Theorem 4.1? A valuable technique we employed in the proof of Theorem 4.1 was using the Gramian to construct an input to transfer the system from the origin to a given state. We extend this technique as follows.

**Corollary 4.3 (State Transfer via Controllability Gramian)** Consider a linear, time-invariant system representation with state equation  $\dot{x}(t) = Ax(t) + Bu(t)$ . If  $(A, B)$  is a controllable pair, then for each time  $T > 0$  and each pair of states  $x_0, x_1 \in \mathbb{R}^n$ , the input  $u : [0, T] \rightarrow \mathbb{R}^m$ ,

$$u(t) = B^\top e^{A^\top(T-t)} W_c(T)^{-1} (x_1 - e^{AT} x_0), \quad (4.41)$$

satisfies  $\varphi(T, 0, x_0, u(\cdot)) = x_1$ .

**Proof** Evaluating the response of the system to the proposed input, we have,

$$\varphi(T, 0, x_0, u(\cdot)) = e^{AT} x_0 + \int_0^T e^{A(T-\tau)} B B^\top e^{A^\top(T-\tau)} W_c(T)^{-1} (x_1 - e^{AT} x_0) d\tau \quad (4.42)$$

$$= e^{AT} x_0 + W_c(T) W_c(T)^{-1} (x_1 - e^{AT} x_0) \quad (4.43)$$

$$= e^{AT} x_0 - e^{AT} x_0 + x_1 \quad (4.44)$$

$$= x_1. \quad (4.45)$$

We conclude that  $u(\cdot)$  satisfies  $\varphi(T, 0, x_0, u(\cdot)) = x_1$ .  $\square$

Thus far, we've shown that the controllability Gramian gives us a *sharp* characterization of controllability and that we can use it to construct a control law that transports the system between any two states. However, we haven't yet developed any method to actually *compute* the controllability Gramian! In the case where  $A$  is Hurwitz, there is a particularly slick method to compute the Gramian using a continuous-time Lyapunov equation.

**Corollary 4.4 (Infinite-Horizon Controllability Gramian via CTLE)** *Consider a linear, time-invariant system with state equation  $\dot{x}(t) = Ax(t) + Bu(t)$ . Suppose  $A$  is Hurwitz. The infinite-horizon controllability Gramian is defined,*

$$W_c = \lim_{t \rightarrow \infty} W_c(t) = \int_0^\infty e^{A\tau} B B^\top e^{A^\top \tau} d\tau. \quad (4.46)$$

$W_c$  satisfies the following two properties:

1.  $W_c$  is the unique solution of the continuous-time Lyapunov equation  $AP + PA^\top = -BB^\top$ .
2.  $W_c$  is positive definite if and only if  $(A, B)$  is a controllable pair.

**Remark 4.12** It's important to remember—this result is only well-posed when  $A$  is Hurwitz! Without this, the infinite-horizon Gramian is *not* guaranteed to converge.

**Remark 4.13** Notice that the Lyapunov equation  $AP + PA^\top = -BB^\top$  uses  $A^\top$  in the place of  $A$ . As we'll see in the proof below, this has no effect on solvability.

**Proof** First, we prove item (1). First, we note that, if  $A$  is Hurwitz, then  $A^\top$  is also Hurwitz. We recall from our earlier study of the continuous-time Lyapunov equation that, when  $A^\top$  is Hurwitz, the unique solution to the CTLE,

$$AP + PA^\top = -BB^\top, \quad (4.47)$$

is given by  $P = \int_0^\infty e^{A\tau} B B^\top e^{A^\top \tau} d\tau$ . We stress that here,  $A$  and  $A^\top$  are flipped compared to their usual order. We immediately recognize this expression as  $\lim_{t \rightarrow \infty} W_c(t)$ . Thus, we conclude that item (1) holds.

Now, we prove item (2). First, we'll show that, if  $W_c$  is positive definite, then  $(A, B)$  is a controllable pair. We'll begin by showing that  $W_c$  being positive definite implies the existence a time  $T$  for which  $W_c(T)$  is positive definite. Suppose  $W_c$  is positive definite.



The minimum singular value of  $W_c$ ,  $\sigma_{\min}(W_c)$ , is a continuous function of  $W_c$  and must be positive if  $W_c$  is positive definite. Thus, one has,

$$\sigma_{\min}(W_c) > 0 \quad (4.48)$$

$$\sigma_{\min}\left(\lim_{t \rightarrow \infty} W_c(t)\right) > 0 \quad (4.49)$$

$$\lim_{t \rightarrow \infty} \sigma_{\min}(W_c(t)) > 0, \quad (4.50)$$

where in the last step, we use continuity of  $\sigma_{\min}$  to extract the limit. If  $\lim_{t \rightarrow \infty} \sigma_{\min}(W_c(t)) > 0$ , there must exist a time  $T > 0$  at which  $\sigma_{\min}(W_c(T)) > 0$ . Since  $W_c(T)$  is always at least positive semidefinite, we conclude that, at time  $T$ ,  $W_c(T)$  is positive definite. By Theorem 4.1, we conclude that the system is controllable.

Now, we show that, if  $(A, B)$  is a controllable pair (and  $A$  is Hurwitz),  $W_c$  is positive definite. If  $(A, B)$  is a controllable pair, then there exists a  $T > 0$  at which  $W_c(T)$  is positive definite. By the definition of  $W_c(T)$ ,  $t \geq T$  implies  $W_c(t) \succeq W_c(T)$ . It follows that  $\lim_{t \rightarrow \infty} W_c(t) \succeq W_c(T) \succ 0$ . We conclude that  $W_c$  is positive definite.  $\square$

This result gives us a nice, algebraic method for determining controllability in the case where  $A$  is Hurwitz—simply solve the CTLE and verify that the solution is positive definite to make a claim of controllability. However, since not every  $A$  matrix will be Hurwitz, this result is clearly insufficient for the general case.

We'd now like to develop a simple method for determining controllability in the general case. Let's begin by identifying where the main problems in computing the Gramian arise. In the formula for the Gramian, we have the integral of a matrix exponential,

$$W_c(T) = \int_0^T e^{A\tau} B B^\top e^{A^\top \tau} d\tau. \quad (4.51)$$

The exponential is arguably the leading factor making  $W_c(T)$  hard to compute. In order to develop a *simpler* characterization of controllability, we'd therefore like a way to reduce the role of the exponential. How can we simplify the matrix exponential in a way that preserves its fundamental algebraic properties? Recall that the exponential is defined by the infinite series,

$$\exp(A) = \sum_{k=0}^{\infty} \frac{A^k}{k!}. \quad (4.52)$$

Can we reduce the information needed to study the exponential in a *lossless* manner? If we could write the higher powers of  $A$  in terms of the lower powers, perhaps we could come up with a substantially simpler way to understand the role of the exponential in controllability. The following, fundamental result from linear algebra gives us exactly this ability.

**Theorem 4.2 (Cayley-Hamilton)** *Consider a matrix  $A \in \mathbb{C}^{n \times n}$  with characteristic polynomial  $\chi_A(s) = s^n + a_{n-1}s^{n-1} + \dots + a_0$ . The matrix  $A$  satisfies,*

$$\chi_A(A) = A^n + a_{n-1}A^{n-1} + \dots + a_0I = 0. \quad (4.53)$$

*Remark 4.14* Since  $\chi_A$  is a polynomial that takes in a single argument  $s \in \mathbb{C}$ , it's not *formally* correct to write  $\chi_A(A)$ . However, since using  $\chi_A(A)$  to refer to the matrix polynomial  $A^n + \dots + a_0I$  is both convenient and unambiguous, it is something we'll often do.

**Proof** We'll prove this result in three stages. First, we'll show that it holds for a Jordan block, then for a matrix in Jordan form, and then for a general matrix  $A \in \mathbb{C}^{n \times n}$ . Let's begin with the case of a Jordan block,  $J_\lambda = \lambda I + N_0 \in \mathbb{C}^{n \times n}$ . Since the only eigenvalue of  $J_\lambda$  is  $\lambda$  and the size of the Jordan block is  $n$ , the characteristic polynomial of  $J_\lambda$  must be,

$$\chi_{J_\lambda}(s) = s^n + a_{n-1}s^{n-1} + \dots + a_0 = (s - \lambda)^n. \quad (4.54)$$

Thus,  $\chi_{J_\lambda}(J_\lambda)$  is computed,

$$\chi_{J_\lambda}(J_\lambda) = J_\lambda^n + a_{n-1}J_\lambda^{n-1} + \dots + a_0I = (J_\lambda - \lambda I)^n = (\lambda I + N_0 - \lambda I)^n = N_0^n = 0, \quad (4.55)$$

where  $N_0^n = 0$  since  $N_0$  is nilpotent of order  $n$ . Thus, the result holds for a Jordan block.

Next, we prove that the result holds for the case of a matrix  $J \in \mathbb{C}^{n \times n}$  in Jordan canonical form. Suppose  $J = \text{blkdiag}(J_{\lambda_1}, \dots, J_{\lambda_k})$ , where each  $J_{\lambda_i} = \lambda_i I + N_0 \in \mathbb{C}^{m_i \times m_i}$  is a Jordan block of size  $m_i$ . The characteristic polynomial of  $J$  is computed,

$$\chi_J(s) = \prod_{i=1}^k (s - \lambda_i)^{m_i}, \quad (4.56)$$

as the product of the characteristic polynomials of each Jordan block. Since  $J$  is a block diagonal matrix and  $\chi_J$  is a polynomial, we have  $\chi_J(J) = \text{blkdiag}(\chi_J(J_{\lambda_1}), \dots, \chi_J(J_{\lambda_k}))$ ,

$$\chi_J(J) = \begin{bmatrix} \prod_{i=1}^k (J_{\lambda_1} - \lambda_i I)^{m_i} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \prod_{i=1}^k (J_{\lambda_k} - \lambda_i I)^{m_i} \end{bmatrix}. \quad (4.57)$$

By the single Jordan block case we proved above, we know that for each  $i \in 1, \dots, k$ ,  $(J_{\lambda_i} - \lambda_i I)^{m_i} = 0$ . Thus, each product on the diagonal of  $\chi_J(J)$  contains a zero, from which we conclude  $\chi_J(J) = 0$ .

Finally, we show the general case. Consider a matrix  $A \in \mathbb{C}^{n \times n}$ . We know there exists a transformation matrix  $T \in \mathbb{C}^{n \times n}$  for which  $T^{-1}AT = J$ , where  $J$  is in Jordan canonical form. Computing the characteristic polynomial of  $J$ , we have,

$$\chi_J(s) = \det(sI - J) = \det(T^{-1}(sI)T - T^{-1}AT) \quad (4.58)$$

$$= \det(T^{-1}) \det(sI - A) \det(T) \quad (4.59)$$

$$= \det(sI - A) = \chi_A(s), \quad (4.60)$$

where we use that  $\det(T^{-1}) = \det(T)^{-1}$  to simplify to  $\det(sI - A)$ .  $J$  and  $A$  therefore have the same characteristic polynomial. Since for any  $k \geq 0$ ,  $(TJT^{-1})^k = TJ^kT^{-1}$ , we have,

$$\chi_A(A) = \chi_A(TJT^{-1}) = \chi_J(TJT^{-1}) = T\chi_J(J)T^{-1}. \quad (4.61)$$

Since  $\chi_J(J) = 0$ , we conclude that  $\chi_A(A) = 0$ . □

The Cayley-Hamilton theorem has a number of incredible consequences for computation of the matrix exponential. We can show that, for a matrix  $A \in \mathbb{C}^{n \times n}$  and any  $k \geq 0$ ,  $A^k$  can be written as a linear combination of the first  $n - 1$  powers of  $A$ . Further, we find that the matrix exponential can be written in terms of the first  $n - 1$  powers of  $A$ .

**Corollary 4.5 (Cayley-Hamilton & Exponentiation)** *Consider a matrix  $A \in \mathbb{R}^{n \times n}$ .*

1. Matrix powers: for any  $k \in \mathbb{Z}_{\geq 0}$ , there exist constants  $\alpha_0, \dots, \alpha_{n-1} \in \mathbb{R}$  for which  $A^k = \alpha_0 I + \dots + \alpha_{n-1} A^{n-1}$ .
2. Matrix exponential: there exist continuously differentiable functions  $\alpha_i \in C^1(\mathbb{R}, \mathbb{R})$ ,  $i = 1, \dots, n-1$  for which  $\exp(At) = \alpha_0(t)I + \dots + \alpha_{n-1}(t)A^{n-1}$ ,  $\forall t \in \mathbb{R}$ .

**Proof** First, we prove item (1). For each of  $k = 0, \dots, n-1$ , the result is trivial (take  $\alpha_k = 1$  and  $\alpha_i = 0$  for  $i \neq k$ ). Now, we prove the case of  $k \geq n$  by induction on  $k$ . First, we prove the base case of  $k = n$ . If  $\chi_A(s) = s^n + a_{n-1}s^{n-1} + \dots + a_0$ , it follows from the Cayley-Hamilton theorem that

$$A^n + a_{n-1}A^{n-1} + \dots + a_0I = 0 \quad (4.62)$$

$$-a_{n-1}A^{n-1} + \dots - a_0I = A^n. \quad (4.63)$$

Thus, we observe that the base case holds. Now, we show the inductive case. Assume for induction that the result holds for some  $k > n$ . Then,

$$A^{k+1} = AA^k = A(\alpha_0 I + \dots + \alpha_{n-1} A^{n-1}) = \alpha_0 A + \dots + \alpha_{n-1} A^n. \quad (4.64)$$

Applying Cayley-Hamilton theorem to  $A^n$ , it follows that,

$$A^{k+1} = \alpha_0 A + \dots + \alpha_{n-1}(-a_{n-1}A^{n-1} + \dots - a_0I). \quad (4.65)$$

Collecting like powers, we conclude that there exist  $\hat{\alpha}_0, \dots, \hat{\alpha}_{n-1} \in \mathbb{R}$  for which  $A^{k+1} = \hat{\alpha}_0 I + \dots + \hat{\alpha}_{n-1} A^{n-1}$ . The inductive case follows, and the result holds for all  $k \in \mathbb{Z}_{\geq 0}$ .

Now, we prove item (2) via an existence and uniqueness argument. In order to show there exist functions  $\alpha_0, \dots, \alpha_{n-1}$  satisfying  $\exp(At) = \sum_{i=0}^{n-1} \alpha_i(t) A^i$ , it is sufficient to show that  $\sum_{i=0}^{n-1} \alpha_i(t) A^i$  solves the matrix initial value problem  $\dot{X}(t) = AX(t)$ ,  $X(0) = I$ . It will then follow by uniqueness of solutions that the sum equals the exponential. Let's write out the details of this argument. Consider an arbitrary collection of differentiable functions  $\alpha_0, \dots, \alpha_{n-1} \in C^1(\mathbb{R}, \mathbb{R})$ . We wish to identify  $\alpha_i$  for which,

$$\frac{d}{dt} \sum_{i=0}^{n-1} \alpha_i(t) A^i = A \sum_{i=0}^{n-1} \alpha_i(t) A^i \quad (4.66)$$

$$\sum_{i=0}^{n-1} \dot{\alpha}_i(t) A^i = \sum_{i=0}^{n-1} \alpha_i(t) A^{i+1}. \quad (4.67)$$

We notice that the final term in the sum on the right hand side contains  $A^n$ . By the Cayley-Hamilton theorem,  $A^n = -a_{n-1}A^{n-1} - \dots - a_0I$ , where each  $a_i$  is a coefficient in the characteristic polynomial of  $A$ . We therefore require the  $\alpha_i$  to satisfy,

$$\sum_{i=0}^{n-1} \dot{\alpha}_i(t) A^i = \sum_{i=0}^{n-2} \alpha_i(t) A^{i+1} - \sum_{i=0}^{n-1} a_i \alpha_{n-1}(t) A^i \quad (4.68)$$

$$= -a_0 \alpha_{n-1}(t) I + \sum_{i=1}^{n-1} (\alpha_{i-1}(t) - a_i \alpha_{n-1}(t)) A^i. \quad (4.69)$$

Now, we equate the coefficients on the left and right hand sides corresponding to the same powers of  $A$ . This results in the set of differential equations,

$$\dot{\alpha}_0(t) = -\alpha_0\alpha_{n-1}(t) \quad (4.70)$$

$$\vdots \quad (4.71)$$

$$\dot{\alpha}_{n-1}(t) = \alpha_{n-1}(t) - a_{n-1}\alpha_{n-1}(t). \quad (4.72)$$

Rewriting these equations in matrix form and applying the initial condition  $\sum_{i=0}^{n-1} \alpha_i(0)A^i = I$ , we find that the components of the solution of the initial value problem,

$$\begin{bmatrix} \dot{\alpha}_0(t) \\ \dot{\alpha}_1(t) \\ \vdots \\ \dot{\alpha}_{n-1}(t) \end{bmatrix} = \begin{bmatrix} 0 & \dots & 0 & -a_0 \\ 1 & \dots & 0 & -a_1 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & 1 & -a_{n-1} \end{bmatrix} \begin{bmatrix} \alpha_0(t) \\ \alpha_1(t) \\ \vdots \\ \alpha_{n-1}(t) \end{bmatrix}, \quad \begin{bmatrix} \alpha_0(0) \\ \alpha_1(0) \\ \vdots \\ \alpha_{n-1}(0) \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad (4.73)$$

are such that  $\sum_{i=0}^{n-1} \alpha_i(t)A^i$  solves the matrix initial value problem  $\dot{X}(t) = AX(t)$ ,  $X(0) = I$ . By uniqueness of solutions to the matrix initial value problem, we conclude

$$\sum_{i=0}^{n-1} \alpha_i(t)A^i = \exp(At). \quad (4.74)$$

Since each  $\alpha_i$  derives from the solution to a linear, time-invariant initial value problem, each  $\alpha_i$  is continuously differentiable. This completes the proof.  $\square$

This result is exactly what we need to eliminate the matrix exponential from our tests for controllability. In particular, it enables us to study controllability through a test on the first  $n-1$  powers of  $A$ .

**Theorem 4.3 (Controllability Matrix)** *Consider a continuous-time, LTI system with state equation  $\dot{x}(t) = Ax(t) + Bu(t)$ ,  $A \in \mathbb{R}^{n \times n}$ . The following are equivalent:*

1. *Controllability:*  $(A, B)$  is a controllable pair.
2. *Controllability matrix:* the controllability matrix,

$$\mathcal{C}_{AB} := [B \ AB \ A^2B \ \dots \ A^{n-1}B], \quad (4.75)$$

*has full row rank (there is no nonzero  $v$  for which  $v^\top \mathcal{C}_{AB} = 0$ ).*

**Proof** First, we show (1)  $\Rightarrow$  (2). Assume that  $(A, B)$  is a controllable pair. We'd like to show that  $\mathcal{C}_{AB}$  has full row rank. Suppose for contradiction that  $\mathcal{C}_{AB}$  does not have full row rank. Then, there exists a vector  $v \in \mathbb{R}^n$  for which,

$$v^\top [B \ AB \ \dots \ A^{n-1}B] = [v^\top B \ v^\top AB \ \dots \ v^\top A^{n-1}B] = 0. \quad (4.76)$$

Thus, for each  $i \in 0, \dots, n-1$ , we have that  $v^\top A^i B = 0$ . From the Cayley-Hamilton theorem, it follows that  $v^\top \exp(At)B = 0$  for all  $t \in \mathbb{R}$ . Thus, for any  $T > 0$ , we have,

$$v^\top W_c(T)v = v^\top \int_0^T e^{A\tau} B B^\top e^{A^\top \tau} d\tau v = \int_0^T v^\top e^{A\tau} B B^\top e^{A^\top \tau} v d\tau = 0, \quad (4.77)$$

which implies that  $W_c(T)$  is not positive definite. But, since we assumed  $(A, B)$  to be a controllable pair, it must be that  $W_c(T) \succ 0$  for all  $T > 0$ . Contradiction! We conclude that  $\mathcal{C}_{AB}$  must have full row rank.

Now, we show that (2)  $\Rightarrow$  (1). Suppose that  $\mathcal{C}_{AB}$  has full row rank. For contradiction, suppose that  $(A, B)$  is not a controllable pair. Then, there exists a  $T > 0$  for which  $W_c(T)$  is not positive definite. Since  $W_c(T)$  is always at least positive semidefinite, this implies that  $W_c(T)$  is singular. Thus, there must exist a nonzero vector  $v \in \mathbb{R}^n$  for which  $v^\top W_c(T)v = 0$ . This implies,

$$\int_0^T \|B^\top e^{A^\top \tau} v\|_2^2 d\tau = 0, \quad (4.78)$$

which in turn implies that  $v^\top e^{At} B = 0$  for all  $t \in [0, T]$ . Thus, for  $i = 1, \dots, n-1$ ,

$$\frac{d^i}{dt^i} v^\top e^{At} B = 0, \quad \forall t \in (0, T) \quad (4.79)$$

$$v^\top A^i e^{At} B = 0, \quad \forall t \in (0, T). \quad (4.80)$$

Taking the limit at  $t \rightarrow 0^+$ , we conclude that  $v^\top A^i B = 0$  for  $i = 1, \dots, n-1$ . Thus,  $v^\top \mathcal{C}_{AB} = 0$ . Since  $v$  is nonzero, this contradicts the assumption that the controllability matrix has full row rank. We conclude that  $W_c(T)$  must be positive definite for all  $T > 0$ , which implies that  $(A, B)$  is a controllable pair.  $\square$

This result gives us a simple, algebraic method for verifying the controllability of a linear, time-invariant system representation. Using this method, all we need to do to check if  $(A, B)$  is a controllable pair is compute the rank of the matrix,

$$\mathcal{C}_{AB} = [B \ AB \ \dots \ A^{n-1}B]. \quad (4.81)$$

Thus, we've achieved our goal of finding a sharp, algebraic test for controllability. Although this test might initially seem like a silver bullet for determining controllability, it's important that we keep its limitations in mind and know where it fits in with respect to the Gramian test for stability.

Let's highlight where the controllability matrix might struggle. Although the controllability matrix *does* give us an easy way to check controllability, it gives no granularity—it is simply a *yes or no* test for controllability. To see why this might be a problem, consider the systems,

$$\dot{x} = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} x + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u \quad (4.82)$$

$$\dot{x} = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} x + \begin{bmatrix} 0 \\ 10^{-7} \end{bmatrix} u. \quad (4.83)$$

Using the controllability matrix test, one may verify that each of the systems is controllable. However, based on the structure of each system, it seems that the first system will be *significantly* easier to control than the second! Since the second system contains a  $10^{-7}$  in its  $B$  matrix, it's *extremely* challenging to control! To enact even minor changes to the state, we would need to use enormous control inputs. The simple yes or no test of controllability offered by the controllability matrix tells us nothing about this disparity.

On the other hand, the controllability Gramian provides us with valuable information about how difficult it is to control each system. By computing the minimum singular value of the controllability Gramian,  $\sigma_{\min}(W_c(T))$ , at some time  $T$ , we can gain a measure of how close the Gramian is to being singular, and therefore how close the system is to being uncontrollable. The closer the minimum singular value is to zero, the harder the system will be to control. Thus, the Gramian—although more difficult to compute—provides more insight into the controllability problem than a simple yes or no test for controllability.

#### 4.1.2.2 The Discrete-Time Case

Now, we perform a similar analysis of controllability for the discrete-time case. We'll find that, for discrete-time systems, almost *identical* characterizations of controllability exist. However, we'll find that we must take more care in specifying the time interval on which controllability holds. Let's get started!

Let's proceed by tracing out the steps we took in the continuous-time case. First, we'll seek out a *Gramian-based* characterization of controllability of a linear, time-invariant state space representation with state equation,

$$x[k+1] = Ax[k] + Bu[k]. \quad (4.84)$$

Recall that, in the continuous-time case, we derived a formula for the controllability Gramian by analyzing the reachability problem on a time interval  $[0, T]$ . Let's do the same for the discrete-time case.

Fix a time  $K \in \mathbb{Z}_{>0}$  and a state  $x \in \mathbb{R}^n$ . We'd like to identify an input  $u[\cdot] : \mathbb{Z} \rightarrow \mathbb{R}^m$  for which  $\varphi(K, 0, 0, u[\cdot]) = x$ . Based on the formula for the discrete-time state transition map, such an input must satisfy,

$$x = \varphi(K, 0, 0, u[\cdot]) = \sum_{j=0}^{K-1} A^{T-j-1} Bu[j]. \quad (4.85)$$

In order to pick an input to *simplify* this sum to a form in which we can perform matrix inversion, we choose the input  $u[j] = B^\top (A^\top)^{T-j-1} Mv$ , where  $M$  and  $v$  are vectors we have yet to choose. Substituting this into the above, we find,

$$\sum_{j=0}^{K-1} A^{K-j-1} Bu[j] = \left[ \sum_{j=0}^{K-1} A^{K-j-1} BB^\top (A^\top)^{K-j-1} \right] Mv \quad (4.86)$$

$$= \left[ \sum_{k=0}^{K-1} A^k BB^\top (A^\top)^k \right] Mv, \quad (4.87)$$

where we performed a change of variables to  $k = K - j - 1$  to clean up the sum. Choosing  $M$  as the inverse of the term in brackets and  $v$  as  $x$  results in an input that delivers us to the state  $x$  at time  $K$ . We conclude that, if the matrix in brackets is invertible, we can solve the reachability (and therefore controllability) problem. This leads to the definition of the discrete-time controllability Gramian.

**Definition 4.8 (Discrete-Time Controllability Gramian)** Consider a discrete-time, LTI system representation  $(A, B, C, D)$ . The discrete-time controllability Gramian is the map  $W_c : \mathbb{N} \rightarrow \mathbb{R}^{n \times n}$ , defined,

$$W_c[K] := \sum_{k=0}^{K-1} A^k B B^\top (A^\top)^k. \quad (4.88)$$

**Exercise 4.4** Confirm that, for any discrete-time, LTI system representation, the discrete-time controllability Gramian is symmetric and positive semidefinite for any  $K \in \mathbb{N}$ .

Using the discrete-time controllability Gramian, we state and prove a discrete-time analogue of Theorem 4.1. It's extremely important to note that, unlike the case of a *continuous-time* system, discrete-time LTI systems *do not* in general enjoy controllability on arbitrarily small time intervals! Rather, the length of the time interval must in general be *at least* equal the size of the  $A$  matrix. We'll discuss this further after proving the theorem.

**Theorem 4.4 (Discrete-Time Gramian Characterization of Controllability)** *Consider a discrete-time, LTI system representation with state equation  $x[k+1] = Ax[k] + Bu[k]$ ,  $A \in \mathbb{R}^{n \times n}$ . The following statements are equivalent:*

1. *There exists a time  $K \in \mathbb{N}$  at which the controllability Gramian  $W_c[K]$  is positive definite.*
2. *For any  $K \in \mathbb{N}_{\geq n}$ , the controllability Gramian is positive definite.*
3. *For any  $K \in \mathbb{N}_{\geq n}$ , the system is reachable on  $[0, K]$ .*
4. *For any  $[k_0, k_1] \subseteq \mathbb{Z}$ ,  $k_1 - k_0 \geq n$ , the system is controllable on  $[k_0, k_1]$ .*

**Proof** We'll prove this result using the same chain of implications as the continuous-time result. We'll show,

$$(1) \Rightarrow (2) \Rightarrow (3) \Rightarrow (4) \Rightarrow (1). \quad (4.89)$$

First, we'll show that  $(1) \Rightarrow (2)$ . Suppose there exists a time  $K \in \mathbb{N}$  at which the controllability Gramian  $W_c[K]$  is positive definite. We'd like to show that the Gramian is positive definite on all of  $\mathbb{N}_{\geq n}$ . For contradiction, suppose there exists a time  $K' \geq n$  at which the Gramian is *not* positive definite. Since the Gramian is always at least positive semidefinite, it follows that the Gramian is singular at  $K'$ . Thus, there exists a  $v \in \mathbb{R}^n$ ,  $v \neq 0$ , for which,

$$0 = v^\top W_c[K'] v = v^\top \sum_{k=0}^{K'-1} A^k B B^\top (A^\top)^k v = \sum_{k=0}^{K'-1} \|B^\top (A^\top)^k v\|_2^2. \quad (4.90)$$

Since the 2-norm is nonnegative, it follows that each term in this sum must equal zero. Since  $K' \geq n$ , we conclude that, for  $k = 0, \dots, n-1$ ,

$$v^\top A^k B = 0. \quad (4.91)$$

By Cayley-Hamilton theorem, we conclude that, for all  $k \geq 0$ ,  $v^\top A^k B = 0$ . It follows that  $v^\top W_c[K] v = 0$ . Contradiction! We conclude that  $W_c[K]$  must be positive definite for all  $K \geq n$ . This completes the proof of  $(1) \Rightarrow (2)$ .

Next, we show that  $(2) \Rightarrow (3)$ . Suppose the controllability Gramian is positive definite for any  $K \geq n$ . Now, fix a particular  $K \geq n$  and a state  $x \in \mathbb{R}^n$ . We'll show that the input,

$$u[k] = B^\top (A^\top)^{K-k-1} W_c[K]^{-1} x, \quad (4.92)$$

which is well-posed due to  $W_c[K] \succ 0$ , drives the system from  $x[0] = 0$  to  $x[K] = x$ . Substituting this input into the formula for the state transition map,

$$\varphi(K, 0, 0, u[\cdot]) = \sum_{j=0}^{K-1} A^{K-j-1} B B^\top (A^\top)^{K-j-1} W_c[K]^{-1} x \quad (4.93)$$

$$= \left[ \sum_{j=0}^{K-1} A^{K-j-1} B B^\top (A^\top)^{K-j-1} \right] W_c[K]^{-1} x \quad (4.94)$$

$$= \left[ \sum_{k=0}^{K-1} A^k B B^\top (A^\top)^k \right] W_c[K]^{-1} x \quad (4.95)$$

$$= W_c[K] W_c[K]^{-1} x = x. \quad (4.96)$$

Thus, we conclude that, for any  $x \in \mathbb{R}^n$ , there exists an input  $u[\cdot]$  satisfying  $\varphi(K, 0, 0, u[\cdot])$ . We conclude that the system is reachable on  $[0, K]$  when  $K \geq n$ .

Next, we show that (3)  $\Rightarrow$  (4). This follows from application of time-invariance. Suppose item (3) holds. Now, fix times  $k_0, k_1 \in \mathbb{Z}$  for which  $k_1 - k_0 \geq N$ . Then, we assume (3), we know that the system is reachable on the interval  $[0, k_1 - k_0]$ . Since the system is LTI, we know that this is equivalent to reachability on  $[k_0, k_1]$ , which in turn is equivalent to controllability on  $[k_0, k_1]$ . We conclude that (3)  $\Rightarrow$  (4).

Finally, we show that (4)  $\Rightarrow$  (1). Suppose the system is controllable on any interval  $[k_0, k_1] \subseteq \mathbb{Z}$ , with  $k_1 - k_0 \geq n$ . We'd like to show there exists a time  $K \in \mathbb{N}$  at which the Gramian is positive definite. Suppose for contradiction that, for all  $K \in \mathbb{N}$ , the Gramian is not positive definite. Fix one such time  $K \geq n$ . If the Gramian  $W_c[K]$  is not positive definite, there exists a vector  $v \in \mathbb{R}^n$ ,  $v \neq 0$  for which,

$$v^\top W_c[K] v = 0 \implies v^\top A^k B = 0, \quad k = 0, \dots, K-1, \quad (4.97)$$

where we use the same reasoning as above to conclude the implication. By assumption, the system is controllable on the interval  $[0, K]$ . Thus, there exists an input signal  $u[\cdot]$  for which  $\varphi(K, 0, 0, u[\cdot]) = v$ . For such a signal, we have,

$$v = \sum_{k=0}^{K-1} A^{K-k-1} B u[k]. \quad (4.98)$$

Left-multiplying both sides by  $v^\top$ , we have,

$$v^\top v = v^\top \sum_{k=0}^{K-1} A^{K-k-1} B u[k] = \sum_{k=0}^{K-1} v^\top A^{K-k-1} B u[k] = 0, \quad (4.99)$$

which implies  $v = 0$ . Since we know  $v \neq 0$ , this is a contradiction. We conclude there exists a time  $K \in \mathbb{N}$  at which the Gramian is positive definite.  $\square$

This theorem has a number of important implications for the control of discrete-time, LTI systems. As we mentioned above, there exists an important difference between the controllability of continuous and discrete-time LTI systems. As opposed to the continuous-time case,



where controllability on one interval implies controllability on any *arbitrarily small* interval, in the discrete-time case, the interval must in general be at least of size  $n$ . To illustrate why this condition is necessary, we consider the following, simple example.

*Example 4.1 (Delay System)* Consider the one-step delay system,

$$\begin{bmatrix} x_1[k+1] \\ x_2[k+1] \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1[k] \\ x_2[k] \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u[k]. \quad (4.100)$$

One may show that this system is controllable. Suppose the system starts at initial condition  $(x_1, x_2) = (0, 1)$ . Can we control the system to the origin in a single time step? A simple analysis reveals that we require *at least two time steps* to drive the system to the origin. To drive the system to the origin in a single time step, we would require the existence of an input  $u[\cdot]$  for which,

$$0 = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u[0] = \begin{bmatrix} 1 \\ u[0] \end{bmatrix}. \quad (4.101)$$

Thus, there is *no* value of  $u[0]$  which takes the system to the origin at time  $k = 1$ . What about at time  $k = 2$ ? We solve,

$$0 = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ u[0] \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u[1] = \begin{bmatrix} u[0] \\ u[1] \end{bmatrix}. \quad (4.102)$$

If we pick  $u[0] = u[1] = 0$ , the system will reach the origin at time  $k = 2$ . We conclude that, for this system, we require at least 2 time steps to transfer between arbitrary states. This matches the dimension of the  $A$  matrix.

As a consequence of this example, we see that *general* discrete-time, LTI systems require a number of time steps equal to the dimension of their  $A$  matrix to reach an arbitrary state.

What are some other consequences of Theorem 4.4? We observe that the controllability of a discrete-time, LTI system on an interval  $[0, K]$ ,  $K \geq n$ , is *entirely* determined by the  $A$  and  $B$  matrices of the system representation. Thus, we make the following definition.

**Definition 4.9 (Discrete-Time Controllable Pair)** A pair of matrices  $(A, B)$ , where  $A \in \mathbb{R}^{n \times n}$  and  $B \in \mathbb{R}^{n \times m}$  is said to be a *discrete-time controllable pair* if there exists an interval on which the system  $x[k+1] = Ax[k] + Bu[k]$  is controllable.

*Remark 4.15* When *discrete-time* is clear from context, we'll refer to a discrete-time controllable pair  $(A, B)$  simply as a controllable pair. Shortly, we'll find that  $(A, B)$  is a discrete-time controllable pair if and only if it's a continuous-time controllable pair! This removes any ambiguity from "controllable pair" referring to the discrete or continuous case.

This directly mirrors the analogous continuous-time definition. What are some other consequences of Theorem 4.4? Recall that, in the continuous-time case, we showed that we can use the controllability Gramian to construct an input that transfers the system between any two states. Now, we do the same for the discrete-time case.

**Corollary 4.6 (Discrete-Time State Transfer via Controllability Gramian)** Consider a discrete-time, LTI system with state equation  $x[k+1] = Ax[k] + Bu[k]$ . If  $(A, B)$  is a controllable pair, then for any  $K \geq n$  and states  $x_0, x_1 \in \mathbb{R}^n$ , the input  $u[\cdot] : [0, K] \rightarrow \mathbb{R}^m$ ,

$$u[k] = B^\top (A^\top)^{K-k-1} W_c[K]^{-1} (x_1 - A^K x_0), \quad (4.103)$$

satisfies  $\varphi(K, 0, x_0, u[\cdot]) = x_1$ .

**Proof** Substituting into the formula for the discrete-time state transition map,

$$\varphi(K, 0, x_0, u[\cdot]) = A^K x_0 + \sum_{j=0}^{K-1} A^{K-j-1} B B^\top (A^\top)^{K-j-1} W_c[K]^{-1} (x_1 - A^K x_0) \quad (4.104)$$

$$= A^K x_0 + \left[ \sum_{j=0}^{K-1} A^{K-j-1} B B^\top (A^\top)^{K-j-1} \right] W_c[K]^{-1} (x_1 - A^K x_0) \quad (4.105)$$

$$= A^K x_0 + W_c[K] W_c[K]^{-1} (x_1 - A^K x_0) \quad (4.106)$$

$$= x_1 + A^K x_0 - A^K x_0 = x_1. \quad (4.107)$$

We conclude that this input signal produces the desired result.  $\square$

Next, we show that in the case where  $A$  is Schur (discrete-time stable), we can compute an infinite-horizon analogue of the discrete-time controllability Gramian using a discrete-time Lyapunov equation.

**Corollary 4.7 (Infinite-Horizon Controllability Gramian via DTLE)** *Consider a discrete-time, LTI system with state equation  $x[k+1] = Ax[k] + Bu[k]$ . Suppose  $A$  is Schur. The infinite-horizon discrete-time controllability gramian is defined,*

$$W_c = \lim_{k \rightarrow \infty} W_c[k] = \sum_{k=0}^{\infty} A^k B B^\top (A^\top)^k. \quad (4.108)$$

$W_c$  satisfies the following two properties:

1.  $W_c$  is the unique solution of the discrete-time Lyapunov equation  $APA^\top - P = -BB^\top$ .
2.  $W_c$  is positive definite if and only if  $(A, B)$  is a controllable pair.

**Remark 4.16** As with the continuous-time case, in the Lyapunov equation used to compute the controllability Gramian, there is an  $A^\top$  in the place of  $A$ ! This is due to the structure of the Gramian as being *mirrored* from the usual solution of the DTLE.

**Exercise 4.5** Prove Corollary 4.7.

Now, we continue to trace the steps we took in the continuous-time case. Recall that, after proving the infinite-horizon Gramian result in the continuous-time case, we identified a simple, algebraic characterization of controllability via the *controllability matrix*. Amazingly, we find that the exact same controllability matrix characterizes the controllability of discrete-time LTI systems.

**Theorem 4.5 (Discrete-Time Controllability Matrix)** *Consider a discrete-time, LTI system with state equation  $x[k+1] = Ax[k] + Bu[k]$ ,  $A \in \mathbb{R}^{n \times n}$ . The following are equivalent:*

1. Controllability:  $(A, B)$  is a controllable pair.
2. Controllability matrix: the controllability matrix,

$$\mathcal{C}_{AB} = [B \ AB \ A^2 B \ \dots \ A^{n-1} B], \quad (4.109)$$

has full row rank (there is no nonzero  $v$  for which  $v^\top \mathcal{C}_{AB} = 0$ ).

**Proof** The proof of this result is facilitated by the following, key observation: the discrete-time controllability Gramian  $W_c[n]$  can be written in terms of the controllability matrix! In particular, the discrete-time controllability Gramian at time  $n$  is calculated,

$$W_c[n] = \sum_{k=0}^{n-1} A^k B B^\top (A^\top)^k = [B \ AB \ \dots \ A^{n-1} B] \begin{bmatrix} B^\top \\ B^\top A^\top \\ \vdots \\ B^\top (A^\top)^{n-1} \end{bmatrix} = \mathcal{C}_{AB} \mathcal{C}_{AB}^\top. \quad (4.110)$$

As a consequence of this equality,  $W_c[n]$  is positive definite if and only if  $\mathcal{C}_{AB}$  is of full row rank. Since  $(A, B)$  is a controllable pair if and only if  $W_c[n]$  is positive definite, we conclude that  $(A, B)$  is a controllable pair if and only if  $\mathcal{C}_{AB}$  has full row rank.  $\square$

### 4.1.3 Decompositions & Stabilizability

What happens when our system *isn't* controllable? Is it always the case that control design is a lost cause? Let's think about when uncontrollability might be “fine.” If all we're interested in doing is stabilizing the system, and the uncontrollable “piece” of the system is already stable, perhaps uncontrollability is fine! This idea leads us to the notion of *stabilizability*, a relaxation of the idea of controllability in which we only request that the “unstable part” of the system be controllable.

In order to formalize the ideas of “controllable part” and “uncontrollable part” system, we must come up with decompositions that transparently split our system into controllable and uncontrollable pieces. How might we do this? By studying the range and kernel of the *controllability matrix*, we can learn a great deal about which parts of the system we can reach and which parts we cannot. Shortly, we'll find that (for sufficiently large  $T$ ), the range of the controllability matrix  $\mathcal{C}_{AB}$  *precisely* equals the reachable subspace  $\mathcal{R}_T$ .

With this in mind, let's try using the controllability matrix to decompose a system into its controllable and uncontrollable “pieces.” First, we state and prove an intermediate lemma. Recall that a subspace  $W \subseteq \mathbb{R}^n$  is  $A$ -invariant for a matrix  $A \in \mathbb{R}^{n \times n}$  if  $v \in W \Rightarrow Av \in W$ . The following result states that the image of the controllability matrix is  $A$ -invariant.

**Lemma 4.1 (Invariance of the Range of  $\mathcal{C}_{AB}$ )** *Consider a pair of matrices  $(A, B)$ ,  $A \in \mathbb{R}^{n \times n}$ ,  $B \in \mathbb{R}^{n \times m}$ . The subspace  $\text{range}(\mathcal{C}_{AB})$  is  $A$ -invariant.*

**Proof** Suppose  $v \in \text{range}(\mathcal{C}_{AB})$ . Then, there exist vectors  $v_0, \dots, v_{n-1}$  for which,

$$v = Bv_0 + ABv_1 + \dots + A^{n-1}Bv_{n-1}. \quad (4.111)$$

Multiplying both sides by  $A$  and applying the Cayley-Hamilton theorem, it follows that there exist vectors  $\hat{v}_0, \dots, \hat{v}_{n-1}$  for which  $Av = \sum_{i=0}^{n-1} A^i B \hat{v}_i$ . We conclude that  $\text{range}(\mathcal{C}_{AB})$  is  $A$ -invariant.  $\square$

Using that  $\text{range}(\mathcal{C}_{AB})$  is  $A$ -invariant, we can come up with a transformation that decomposes a system into its controllable and uncontrollable components.

**Theorem 4.6 (Controllability Decomposition)** *Consider a continuous/discrete time LTI system representation  $(A, B, C, D)$ , for which  $\dim(\text{range}(\mathcal{C}_{AB})) = r \leq n$ . There exists an invertible matrix  $T \in \mathbb{R}^{n \times n}$  for which,*

$$T^{-1}AT = \begin{bmatrix} \hat{A}_c & \hat{A}_{12} \\ 0 & \hat{A}_{uc} \end{bmatrix}, \quad T^{-1}B = \begin{bmatrix} \hat{B}_c \\ 0 \end{bmatrix}, \quad (4.112)$$

where  $\hat{A}_c \in \mathbb{R}^{r \times r}$  and  $\hat{B}_c \in \mathbb{R}^{r \times m}$  form a controllable pair.

*Remark 4.17* A system with  $A$  and  $B$  matrices of the form

$$\begin{bmatrix} \hat{A}_c & \hat{A}_{12} \\ 0 & \hat{A}_{uc} \end{bmatrix}, \quad \begin{bmatrix} \hat{B}_c \\ 0 \end{bmatrix}, \quad (4.113)$$

where  $(\hat{A}_c, \hat{B}_c)$  is a controllable pair, is said to be in *controllability form*.

*Remark 4.18* The controllability form of a system representation is *not* in general unique! In the proof below, we'll see that a number of controllability forms can be constructed for a single representation.

**Proof** We provide a constructive proof of this result. First, select a basis  $\{v_1, \dots, v_r\}$  for  $\text{range}(\mathcal{C}_{AB})$  and a basis  $\{v_{r+1}, \dots, v_n\}$  for its orthogonal complement,  $\text{range}(\mathcal{C}_{AB})^\perp$ . Since  $\text{range}(\mathcal{C}_{AB})$  is  $A$ -invariant, it follows that, for each  $v_i, i = 1, \dots, r$ , there exist constants  $c_1, \dots, c_r$  for which

$$Av_i = c_1 v_1 + \dots + c_r v_r. \quad (4.114)$$

Thus, there exists a matrix  $\hat{A}_c \in \mathbb{R}^{r \times r}$  for which

$$A \begin{bmatrix} | & & | \\ v_1 & \dots & v_r \\ | & & | \end{bmatrix} = \begin{bmatrix} | & & | \\ v_1 & \dots & v_r \\ | & & | \end{bmatrix} \hat{A}_c. \quad (4.115)$$

The product  $AT$  therefore satisfies,

$$AT = A \begin{bmatrix} | & & | & & | \\ v_1 & \dots & v_r & \dots & v_n \\ | & & | & & | \end{bmatrix} = T \begin{bmatrix} \hat{A}_c & \hat{A}_{12} \\ 0 & \hat{A}_{uc} \end{bmatrix}, \quad (4.116)$$

for some matrices  $\hat{A}_{12}, \hat{A}_{uc}$ , where  $\hat{A}_c \in \mathbb{R}^{r \times r}$ . Since  $\{v_1, \dots, v_r, \dots, v_n\}$  is a basis for  $\mathbb{R}^n$ ,  $T$  is invertible. We conclude that

$$T^{-1}AT = \begin{bmatrix} \hat{A}_c & \hat{A}_{12} \\ 0 & \hat{A}_{uc} \end{bmatrix}. \quad (4.117)$$

Now, we show that  $T^{-1}B$  has the desired structure. Since each column of  $B$  is in the range of the controllability matrix, it follows that each column of  $B$  can be written as a linear combination of  $v_1, \dots, v_r$ . Thus, there exists a matrix  $\hat{B}_c$  for which

$$B = [v_1 \dots v_r] \hat{B}_c = T \begin{bmatrix} \hat{B}_c \\ 0 \end{bmatrix}. \quad (4.118)$$

Thus,  $T^{-1}B$  has the desired form. Now, we show that  $(\hat{A}_c, \hat{B}_c)$  are a controllable pair. We begin by computing the controllability matrix of the transformed system. We have,

$$[T^{-1}B \mid T^{-1}AT \cdot T^{-1}B \mid \dots \mid T^{-1}A^{n-1}T \cdot T^{-1}B] = \begin{bmatrix} \hat{B}_c & \hat{A}_c \hat{B}_c & \dots & \hat{A}_c^{n-1} \hat{B}_c \\ 0 & 0 & \dots & 0 \end{bmatrix} \quad (4.119)$$

$$T^{-1} [B \ AB \ \dots \ A^{n-1}B] = \begin{bmatrix} \hat{B}_c & \hat{A}_c \hat{B}_c & \dots & \hat{A}_c^{n-1} \hat{B}_c \\ 0 & 0 & \dots & 0 \end{bmatrix} \quad (4.120)$$

$$T^{-1}\mathcal{C}_{AB} = \begin{bmatrix} \hat{B}_c & \hat{A}_c \hat{B}_c & \dots & \hat{A}_c^{n-1} \hat{B}_c \\ 0 & 0 & \dots & 0 \end{bmatrix}. \quad (4.121)$$

Since  $T^{-1}$  is invertible, both the left and right hand sides have equal rank. Since  $\hat{A}_c \in \mathbb{R}^{r \times r}$ , we may apply the Cayley-Hamilton theorem to conclude,

$$\text{rank} [\hat{B}_c \ \hat{A}_c \hat{B}_c \ \dots \ \hat{A}_c^{r-1} \hat{B}_c] = r. \quad (4.122)$$

Thus,  $\mathcal{C}_{\hat{A}_c \hat{B}_c}$  is full rank. We conclude that  $(\hat{A}_c, \hat{B}_c)$  is a controllable pair.  $\square$

**Exercise 4.6** Explain why the controllability form of a general, LTI system is non-unique. *Hint: can you change the order of the columns of  $T$  without breaking the structure?*

Let's interpret this result in the case of a continuous-time system. If we transform a continuous-time system into its controllability form, we'll find,

$$\begin{bmatrix} \dot{x}_c \\ \dot{x}_{uc} \end{bmatrix} = \begin{bmatrix} \hat{A}_c & \hat{A}_{12} \\ 0 & \hat{A}_{uc} \end{bmatrix} \begin{bmatrix} x_c \\ x_{uc} \end{bmatrix} + \begin{bmatrix} \hat{B}_c \\ 0 \end{bmatrix} u. \quad (4.123)$$

From this decomposition, we observe that the controllability form makes it *extremely transparent* to determine which parts of the system we can control and which parts we cannot. Notably, the uncontrollable component, which we denote above as  $x_{uc}$ , doesn't even have an input showing up in its dynamics! This makes it clear that we have *no control* over the evolution of  $x_{uc}$ .

How does the controllability decomposition interact with the *reachable subspace* we defined earlier in the section? Recall that, earlier, we defined the reachable subspace,

$$\mathcal{R}_T = \{x \in \Sigma : (0, 0) \rightsquigarrow (x, T)\}, \quad (4.124)$$

as the set of all points we could reach from the origin in time  $T > 0$ . Now, we confirm that the reachable subspace of the system *exactly* equals  $\text{range}(\mathcal{C}_{AB})$  (provided  $T$  is sufficiently large in the discrete-time case). This result confirms that the controllability decomposition preserves the structure of the reachable subspace.

**Corollary 4.8 (Reachable Subspace Equals the Range of  $\mathcal{C}_{AB}$ )** Consider a linear, time-invariant system representation  $(A, B, C, D)$ ,  $A \in \mathbb{R}^{n \times n}$ .

1. *Continuous-time:* if the system is continuous-time, then  $\forall T \in \mathbb{R}_{>0}$ ,  $\mathcal{R}_T = \text{range}(\mathcal{C}_{AB})$ .
2. *Discrete-time:* if the system is discrete-time, then  $\forall K \in \mathbb{N}_{\geq n}$ ,  $\mathcal{R}_K = \text{range}(\mathcal{C}_{AB})$ .

*Remark 4.19* Below, we'll give an "easy" proof of this result using the controllability decomposition. There is a more direct but challenging proof, which doesn't rely on this decomposition, that involves use of the controllability Gramian. For a proof using this method, you may consult [14], Lemmas 2.6 and 2.8.

**Proof** First, we'll prove the continuous-time case. Consider a continuous-time, LTI system representation  $(A, B, C, D)$ ,  $A \in \mathbb{R}^{n \times n}$ , with controllability matrix  $\mathcal{C}_{AB}$  of rank  $r \leq n$ . By Theorem 4.6, we can transform this system into the form,

$$\begin{bmatrix} \dot{x}_c \\ \dot{x}_{uc} \end{bmatrix} = \begin{bmatrix} \hat{A}_c & \hat{A}_{12} \\ 0 & \hat{A}_{uc} \end{bmatrix} \begin{bmatrix} x_c \\ x_{uc} \end{bmatrix} + \begin{bmatrix} \hat{B}_c \\ 0 \end{bmatrix} u. \quad (4.125)$$

Here,  $(\hat{A}_c, \hat{B}_c)$  is controllable and  $(x_c, u_c) = Tx$ , where  $T = [v_1, \dots, v_r, \dots, v_n]$  is formed from a basis for  $\text{range}(\mathcal{C}_{AB})$  and its orthogonal complement.

Fix a time  $T > 0$  and a vector  $x \in \text{range}(\mathcal{C}_{AB})$ . We will show  $x \in \mathcal{R}_T$ . If  $x \in \text{range}(\mathcal{C}_{AB})$ ,  $Tx$  is of the form  $(x_c, 0)$ , where  $x_c \in \mathbb{R}^r$ . Since  $(\hat{A}_c, \hat{B}_c)$  is controllable, there exists an input  $\hat{u}(\cdot)$  for which the solution  $x_c(t)$  to

$$\dot{x}_c(t) = \hat{A}_c x_c(t) + \hat{B}_c \hat{u}(t), \quad x_c(0) = 0, \quad (4.126)$$

satisfies  $x_c(T) = x_c$ . On the other hand, the solution to the uncontrollable initial value problem  $\dot{x}_{uc}(t) = \hat{A}_{uc} x_{uc}(t)$ ,  $x_{uc}(0) = 0$  is identically zero. Thus, we conclude that the solution to the initial value problem,

$$\begin{bmatrix} \dot{x}_c \\ \dot{x}_{uc} \end{bmatrix} = \begin{bmatrix} \hat{A}_c & \hat{A}_{12} \\ 0 & \hat{A}_{uc} \end{bmatrix} \begin{bmatrix} x_c \\ x_{uc} \end{bmatrix} + \begin{bmatrix} \hat{B}_c \\ 0 \end{bmatrix} \hat{u}(t), \quad \begin{bmatrix} x_c(0) \\ x_{uc}(0) \end{bmatrix} = 0, \quad (4.127)$$

satisfies  $(x_c(T), 0) = (x_c, 0) = Tx$ . The same input  $\hat{u}(t)$  therefore satisfies  $x = \varphi(T, 0, 0, u(\cdot))$  for our original system. Thus,  $x \in \mathcal{R}_T$ , and  $\text{range}(\mathcal{C}_{AB}) \subseteq \mathcal{R}_T$ .

Now, we prove the reverse direction of inclusion. Fix a time  $T > 0$  and a vector  $x \in \mathcal{R}_T$ . We wish to show that  $x \in \text{range}(\mathcal{C}_{AB})$ . This is immediate from the controllability decomposition. We conclude that  $\mathcal{R}_T \subseteq \text{range}(\mathcal{C}_{AB})$ . Together with the above, this implies that  $\mathcal{R}_T = \text{range}(\mathcal{C}_{AB})$ . The discrete-time case follows from an identical argument.  $\square$

**Exercise 4.7** Complete the proof of Corollary 4.8 by proving the discrete-time case. In general, what is the minimum  $K$  for which  $\text{range}(\mathcal{C}_{AB}) = \mathcal{R}_K$ ?

As a consequence of this result, we could just as well take  $\text{range}(\mathcal{C}_{AB})$  to be the definition of the reachable subspace  $\mathcal{R}_T$  (provided  $T \geq n$  for the discrete-time case). Due to this equivalence, we find that the reachable subspace is also  $A$ -invariant.

**Exercise 4.8** Using Corollary 4.8 and Lemma 4.1, confirm that the reachable subspace  $\mathcal{R}_T$  of a continuous-time system is  $A$ -invariant for all  $T > 0$ . What can you say about the discrete-time case?

Finally, we return to our original motivation for studying the controllability decomposition: *stabilizability*. Now that we've come up with a way to split a system into its controllable and uncontrollable pieces, we can declare what it means for the uncontrollable part of the system to be stable.

**Definition 4.10 (Stabilizability)** Consider a continuous/discrete-time LTI system representation  $(A, B, C, D)$  with a controllability form,

$$\begin{bmatrix} \hat{A}_c & \hat{A}_{12} \\ 0 & \hat{A}_{uc} \end{bmatrix}, \quad \begin{bmatrix} \hat{B}_c \\ 0 \end{bmatrix}. \quad (4.128)$$

Stabilizability is defined in the continuous and discrete-time cases as follows:

1. Continuous-time: the system is stabilizable if  $\hat{A}_{uc}$  is Hurwitz.
2. Discrete-time: the system is stabilizable if  $\hat{A}_{uc}$  is Schur.

If the system is stabilizable, we say that  $(A, B)$  is a *stabilizable pair*.

*Remark 4.20* Recall that the controllability form of an arbitrary LTI system is not guaranteed to be unique! As such, to ensure that Definition 4.10 is well-posed, we must check that stabilizability is a property of the *system* and not a property of a particular controllability form. The details of this are requested in the exercise below.

**Exercise 4.9** Confirm that the definition of stabilizability is well-posed by checking that, if one controllability form of the system satisfies the conditions for stabilizability, then every controllability form must satisfy the conditions for stabilizability.

Thus, a system is *stabilizable* if its uncontrollable component is naturally exponentially stable. If we wish to design a controller to stabilize such a system, all we need to do is focus on the controllable component—the uncontrollable component already has the desired stability properties.

Definition 4.10 suggests one potential method for verifying the stabilizability of a system—transform into controllability form and calculate the eigenvalues of  $\hat{A}_{uc}$  to check if it is Hurwitz or Schur. The *Popov-Belevich-Hautus (PBH)* tests provide more direct tests for stabilizability, as well as another equivalent test for controllability.

**Theorem 4.7 (PBH Tests)** Consider a continuous/discrete-time LTI system representation  $(A, B, C, D)$ . The following tests for controllability and stabilizability hold.

1. Controllability:  $(A, B)$  is a controllable pair if and only if the matrix  $[A - \lambda I, B]$  has full row rank for all  $\lambda \in \mathbb{C}$ .
2. Continuous-time stabilizability:  $(A, B)$  is a continuous-time stabilizable pair if and only if the matrix  $[A - \lambda I, B]$  has full row rank for all  $\lambda \in \mathbb{C}$  such that  $\text{Re}(\lambda) \geq 0$ .
3. Discrete-time stabilizability:  $(A, B)$  is a discrete-time stabilizable pair if and only if the matrix  $[A - \lambda I, B]$  has full row rank for all  $\lambda \in \mathbb{C}$  such that  $|\lambda| \geq 1$ .

**Proof** First, we'll prove item (1). Suppose  $(A, B)$  is a controllable pair. For contradiction, suppose there exists a  $\lambda \in \mathbb{C}$  for which the matrix  $[A - \lambda I, B]$  does not have full row rank. Then, there exists a nonzero vector  $v \in \mathbb{C}^n$  for which,

$$v^* [A - \lambda I \ B] = 0 \quad (4.129)$$

$$[v^* (A - \lambda I) \ v^* B] = 0. \quad (4.130)$$

Thus,  $v^*$  satisfies  $v^* A = \lambda v^*$  and  $v^* B = 0$ . Left-multiplying the controllability matrix  $\mathcal{C}_{AB}$  by  $v^*$ , it follows that,

$$v^* [B \ AB \ \dots \ A^{n-1}B] = [v^* B \ \lambda v^* B \ \dots \ \lambda^{n-1} v^* B] = 0. \quad (4.131)$$

Since  $v$  is nonzero,  $\mathcal{C}_{AB}$  cannot have full row rank. Since we assumed  $(A, B)$  to be a controllable pair, this is a contradiction! This completes the forward direction.

Now, we prove the reverse direction. Suppose  $[A - \lambda I, B]$  has full row rank for all  $\lambda \in \mathbb{C}$ . For contradiction, suppose  $(A, B)$  is *not* a controllable pair. Then, there exists a transformation  $T$  for which,

$$T^{-1}AT = \begin{bmatrix} \hat{A}_c & \hat{A}_{12} \\ 0 & \hat{A}_{uc} \end{bmatrix}, \quad T^{-1}B = \begin{bmatrix} \hat{B}_c \\ 0 \end{bmatrix}. \quad (4.132)$$

Now, pick any eigenvalue  $\lambda_{uc}$  corresponding to the uncontrollable block  $\hat{A}_{uc}$  and a corresponding left eigenvector  $v_{uc}$  of  $\hat{A}_{uc}$ . Defining the vector  $v = (0, v_{uc})$ , it follows that,

$$v^* [T^{-1}AT - \lambda_{uc}I \mid T^{-1}B] = 0. \quad (4.133)$$

Thus, the matrix  $[T^{-1}AT - \lambda_{uc}I \mid T^{-1}B]$  is not of full row rank. Factoring out the  $T^{-1}$  and  $T$  terms, we find,

$$[T^{-1}AT - \lambda_{uc}I \mid T^{-1}B] = T^{-1} [A - \lambda_{uc}I \mid B] \begin{bmatrix} T & 0 \\ 0 & I \end{bmatrix}. \quad (4.134)$$

Since the matrices on the left and right hand side of  $[A - \lambda_{uc}I, B]$  are nonsingular, the row rank of  $[A - \lambda_{uc}I, B]$  must equal that of  $[T^{-1}AT - \lambda_{uc}I, T^{-1}B]$ . Thus,  $[A - \lambda_{uc}I, B]$  is not full row rank. As this is a contradiction, we conclude that the reverse direction holds.

Items (2) and (3) follow from similar reasoning to item (1) and from the observation that the matrix  $[A - \lambda I, B]$  drops rank when  $\lambda$  is an eigenvalue of  $A$ . If the matrix is full rank for all unstable eigenvalues, it follows that the unstable component of the system must be controllable.  $\square$

**Exercise 4.10** Fill in the details of the proofs of items (2) and (3).

The sharp characterization of stabilizability given by the PBH test motivates the following definition.

**Definition 4.11 (Uncontrollable Mode)** Consider a continuous/discrete-time LTI system representation  $(A, B, C, D)$ ,  $A \in \mathbb{R}^{n \times n}$ . A value  $\lambda \in \mathbb{C}$  is said to be an *uncontrollable mode* of the system if

$$\text{rank} [A - \lambda I \mid B] < n. \quad (4.135)$$

In this language, one may alternatively phrase the definition of stabilizability of a system as the property, “all uncontrollable modes are stable.”

#### 4.1.4 Minimum-Energy Control

Let’s summarize what we’ve done so far. We began the section by outlining a general framework for studying controllability. We then focused on the case of linear, time-invariant systems, and developed Gramian-based tests for controllability. Following this, we derived a simpler algebraic test for controllability with the controllability matrix. Then, we discussed applications of the controllability matrix in decomposing systems into controllable and uncontrollable pieces. We then found the range of a system’s controllability matrix to be exactly equal to its reachable subspace.

We finish our study of controllability by returning to the Gramian-based characterization we introduced all the way at the beginning of the section. Recall that, after motivating the Gramian, we came up with two “magic” input formulas,



$$u(t) = B^\top e^{A^\top(T-t)} W_c(T)^{-1} (x_1 - e^{AT} x_0) \quad (\text{Continuous-time}) \quad (4.136)$$

$$u[k] = B^\top (A^\top)^{K-k-1} W_c[K]^{-1} (x_1 - A^K x_0) \quad (\text{Discrete-time}), \quad (4.137)$$

which transported the system between the states  $x_0, x_1 \in \mathbb{R}^n$  across the intervals  $[0, T]$  and  $[0, K]$  (provided  $K \geq n$ ). Does this input formula have any significance beyond its utility in the Gramian characterization of controllability?

Let's take a second look at the setup of the controllability problem to answer this question. In continuous-time, when deriving the Gramian, we found ourselves looking for an input  $u(\cdot)$  satisfying the equality,

$$x = \int_0^T e^{A(T-\tau)} B u(\tau) d\tau. \quad (4.138)$$

How else can we view this equality? Let's think of the integral term as an *operator*,  $L$ , from the input space  $\mathcal{U}$  to the state space  $\mathbb{R}^n$ .

$$L : \mathcal{U} \rightarrow \mathbb{R}^n \quad (4.139)$$

$$u(\cdot) \mapsto Lu = \int_0^T e^{A(T-\tau)} B u(\tau) d\tau. \quad (4.140)$$

Using this operator, the problem we were trying to solve when deriving the Gramian was,

$$x = Lu \quad (4.141)$$

Let's pretend for a moment that  $L$  is a matrix in  $\mathbb{R}^{n \times m}$  and  $u$  is a vector in  $\mathbb{R}^m$ . In this context, one tool we can use to solve the problem  $x = Lu$  is *Moore-Penrose pseudoinverse* of  $L$ , which, provided  $LL^\top$  is nonsingular, is defined  $L^+ = L^\top(LL^\top)^{-1}$ . In the event where  $LL^\top$  is nonsingular, the vector  $u = L^\top(LL^\top)^{-1}x$  solves the desired problem, since

$$Lu = L^\top(LL^\top)^{-1}x = x. \quad (4.142)$$

What's more, one can show that the choice of  $u = L^\top(LL^\top)^{-1}x$  is the smallest  $\ell_2$ -norm vector solving  $x = Lu$ . That is,  $u = L^\top(LL^\top)^{-1}x$  is the unique solution of the minimization problem,

$$\inf_u \|u\|_2^2, \quad \text{subject to } x = Lu. \quad (4.143)$$

Let's return from the simple matrix-vector problem to controllability problem setting, where  $L : \mathcal{U} \rightarrow \mathbb{R}^n$  is the linear operator  $Lu = \int_0^T e^{A(T-\tau)} B u(\tau) d\tau$  from the function space  $\mathcal{U}$  to  $\mathbb{R}^n$ . Can we solve the problem  $x = Lu$  using a functional analytic analogue of the Moore-Penrose pseudoinverse? If so, does such a solution still possess a minimum norm property?

Amazingly, the answer to both of these questions is *yes*. In particular, we'll find that the Gramian input formulas we posed at the beginning of the section are *exactly* derived from this functional analytic analogue of the Moore-Penrose pseudoinverse.

**Theorem 4.8 (Gramian Inputs Minimize Energy)** *Consider a continuous/discrete-time LTI system representation  $(A, B, C, D)$ ,  $A \in \mathbb{R}^{n \times n}$ , for which  $(A, B)$  is a controllable pair. Fix a pair of states  $x_0, x_1 \in \mathbb{R}^n$ .*

1. *Continuous-time:* for any  $T > 0$ , the input  $u(t) = B^\top e^{A^\top(T-t)} W_c(T)^{-1} (x_1 - e^{AT} x_0)$  is the unique input achieving the infimum,

$$\inf_{u(\cdot) \in L^2([0, T], \mathbb{R}^m)} \int_0^T \|u(\tau)\|_2^2 d\tau, \text{ subject to } \varphi(T, 0, x_0, u(\cdot)) = x_1. \quad (4.144)$$

2. *Discrete-time:* for any  $K \geq n$ , the input  $u[k] = B^\top (A^\top)^{K-k-1} W_c[K]^{-1} (x_1 - A^K x_0)$  is the unique input achieving the infimum,

$$\inf_{u[\cdot]: [0, K-1] \rightarrow \mathbb{R}^m} \sum_{k=0}^{K-1} \|u(k)\|_2^2 d\tau, \text{ subject to } \varphi(K, 0, x_0, u[\cdot]) = x_1. \quad (4.145)$$

*Remark 4.21* In each of the minimization problems posed above, we seek to minimize the square of the 2-norm of a signal across the interval  $[0, T]$ . Since the square of the 2-norm is often interpreted as the energy of the signal (as we discussed in the previous chapter), we say that the Gramian inputs above *minimize energy*.

In order to prove Theorem 4.8, we require a few intermediate results on the pseudoinverse in infinite-dimensional inner product spaces. After stating and proving these results, we'll return to prove Theorem 4.8 in the continuous and discrete-time cases.

#### 4.1.4.1 The Moore-Penrose Pseudoinverse ★

In order to generalize the Moore-Penrose pseudoinverse to the setting of abstract function spaces, we first need to introduce a class of well-behaved inner product spaces. We recall the definition of an inner product space below.

**Definition 4.12 (Inner Product Space)** Let  $V$  be a vector space over a field  $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$ . An inner product on  $V$  is a map  $\langle \cdot, \cdot \rangle : V \times V \rightarrow \mathbb{K}$  satisfying:

1. Positive definite:  $\langle v, v \rangle \geq 0$  for all  $v \in V$  and  $\langle v, v \rangle = 0 \Leftrightarrow v = 0$ .
2. Conjugation:  $\langle v, w \rangle = \overline{\langle w, v \rangle}$ , for all  $v, w \in V$ .
3. Linearity:  $\langle u, \alpha v + \beta w \rangle = \alpha \langle u, v \rangle + \beta \langle u, w \rangle$ , for all  $u, v, w \in V$  and  $\alpha \in \mathbb{K}$ .

A pair of a vector space  $V$  and an inner product  $\langle \cdot, \cdot \rangle$  on  $V$  is called an *inner product space*.

One may verify that an inner product induces a natural choice of norm on a vector space. In particular, given an inner product space  $(V, \langle \cdot, \cdot \rangle)$ , one may define a norm on  $V$  by,

$$\|v\| = \sqrt{\langle v, v \rangle}. \quad (4.146)$$

This particular norm is said to be *induced by the inner product*  $\langle \cdot, \cdot \rangle$ . An inner product space in which the norm  $\|\cdot\|$  induced by the inner product makes  $(V, \|\cdot\|)$  a *complete normed vector space* is called a *Hilbert space*.

**Definition 4.13 (Hilbert Space)** Let  $(V, \langle \cdot, \cdot \rangle)$  be an inner product space and  $\|\cdot\|$  be the norm induced by  $\langle \cdot, \cdot \rangle$ . The space  $(V, \langle \cdot, \cdot \rangle)$  is said to be a *Hilbert space* if  $(V, \|\cdot\|)$  is a complete normed vector space.

*Remark 4.22* As a consequence of this definition, the set of Hilbert space is in fact a *subset* of the set of Banach spaces.

An example of a Hilbert space which we've already encountered in this course is the space  $L^2(\mathbb{R}, \mathbb{R}^m)$  of square-integrable functions from  $\mathbb{R}$  to  $\mathbb{R}^m$ . Let's identify the inner product for this Hilbert space.

*Example 4.2 ( $L^2$ -space)* Recall that, when studying norms on signals and systems, we defined the space  $L^2(\mathbb{R}, \mathbb{R}^m)$  to be the set of all *square-integrable* functions on  $\mathbb{R}$  with values in  $\mathbb{R}^m$ . That is,  $L^2(\mathbb{R}, \mathbb{R}^m)$  is the set of functions  $u(\cdot) : \mathbb{R} \rightarrow \mathbb{R}^m$  for which,

$$\|u\|_2^2 = \int_{\mathbb{R}} \|u(\tau)\|_2^2 d\tau < \infty. \quad (4.147)$$

Let's show that the  $L^2$  norm is induced by an inner product. We define the  $L^2$  inner product,

$$\langle u, v \rangle_{L^2} = \int_{\mathbb{R}} u(\tau)^\top v(\tau) d\tau, \quad (4.148)$$

for functions  $u(\cdot), v(\cdot) : \mathbb{R} \rightarrow \mathbb{R}^m$ . Based on the definition of the  $L^2$  norm, it's clear that  $\|u\|_2^2 = \langle u, u \rangle_{L^2}$ . One can show  $L^2(\mathbb{R}, \mathbb{R}^m)$  is a Hilbert space with the inner product  $\langle \cdot, \cdot \rangle_{L^2}$ .

This example is easily generalized from functions with a domain of  $\mathbb{R}$  to functions with a domain of a compact interval  $I \subseteq \mathbb{R}$ . For instance, we can make the space  $L^2([0, T], \mathbb{R}^m)$ ,  $T > 0$  of square-integrable functions on  $[0, T]$  into a Hilbert space by defining the inner product,

$$\langle u, v \rangle = \int_0^T u(\tau)^\top v(\tau) d\tau. \quad (4.149)$$

We'll focus on this Hilbert space in our study of the continuous-time controllability Gramian.

*Remark 4.23* As in the case of the normed vector space  $L^2(\mathbb{R}, \mathbb{R}^m)$ , we make a brief disclaimer regarding the definition of the inner product space  $(L^2(\mathbb{R}, \mathbb{R}^m), \langle \cdot, \cdot \rangle_{L^2})$  defined above. If we take the elements of  $L^2(\mathbb{R}, \mathbb{R}^m)$  to be *functions*  $u : \mathbb{R} \rightarrow \mathbb{R}^m$ , the inner product posed above is *not* a formal inner product for the same reason that the  $L^2$  norm is not a formal norm. If we take a function  $u : \mathbb{R} \rightarrow \mathbb{R}^m$  which is zero except on a nonempty set of measure zero,  $u$  will satisfy  $\langle u, u \rangle_{L^2} = 0$  despite not being the zero function. Because of this, the  $L^2$  "inner product" we defined above is actually a *pseudo* inner product. One can rectify this issue and make  $L^2(\mathbb{R}, \mathbb{R}^m)$  into a true inner product space by taking its elements to be sets of functions that are equal everywhere except on sets of measure zero. As with the case of the normed vector space  $L^2(\mathbb{R}, \mathbb{R}^m)$  and the  $L^2$ -norm, we'll ignore this technical detail in our treatment of the material and treat  $L^2(\mathbb{R}, \mathbb{R}^m)$  as a Hilbert space whose elements are functions.

Since Hilbert spaces can be thought of as complete normed vector spaces, all of the standard definitions of induced norms and bounded linear operators in normed vector spaces carry over to the Hilbert space setting. Using the notion of an *adjoint*, one generalizes the standard matrix transpose to arbitrary Hilbert spaces.

**Definition 4.14 (Adjoint)** Consider a pair of Hilbert spaces  $(H_1, \langle \cdot, \cdot \rangle_1), (H_2, \langle \cdot, \cdot \rangle_2)$  over a common field  $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$  and a bounded linear operator  $L : H_1 \rightarrow H_2$ . The *adjoint* of  $L$  is the bounded linear operator  $L^* : H_2 \rightarrow H_1$  satisfying

$$\langle Lu, v \rangle_2 = \langle u, L^*v \rangle_1, \quad \forall u \in H_1, v \in H_2. \quad (4.150)$$

It is a deep theorem from functional analysis that a unique adjoint exists for any bounded linear operator between two Hilbert spaces over a common field. To get a feel for this definition, let's try computing the adjoint of a special linear operator on  $L^2([0, T], \mathbb{R}^m)$ .

*Example 4.3 (Adjoint on  $L^2$ )* Fix a value  $T > 0$ . Let's define a linear operator with a domain of  $L^2([0, T], \mathbb{R}^m)$  and a codomain of  $\mathbb{R}^n$ . Let  $M : [0, T] \rightarrow \mathbb{R}^{m \times n}$  be a matrix-valued function whose entries belong to  $L_2([0, T], \mathbb{R})$ . For any  $u \in L^2([0, T], \mathbb{R}^m)$ , define  $L_M u$  as,

$$L_M u = \int_0^T M(\tau)^\top u(\tau) d\tau. \quad (4.151)$$

Let's compute the adjoint of  $L_M$  as an operator from  $(L^2([0, T], \mathbb{R}^m), \langle \cdot, \cdot \rangle_{L^2})$  to  $(\mathbb{R}^n, \langle \cdot, \cdot \rangle_2)$ ,  $\mathbb{R}^n$  with the standard Euclidean inner product (one may show that this is also a Hilbert space). Since for any  $x \in \mathbb{R}^n$  the adjoint must satisfy  $\langle L_M u, x \rangle_2 = \langle u, L_M^* x \rangle_2$ , we must have,

$$\left( \int_0^T M(\tau)^\top u(\tau) d\tau \right)^\top x = \int_0^T u^\top(\tau) (L_M^* x)(\tau) d\tau. \quad (4.152)$$

From this equality, it follows that  $(L_M^* x)(t) = M(t)x$ . Thus, the adjoint maps a vector  $x \in \mathbb{R}^n$  to the function  $M(\cdot)x$  in  $L^2([0, T], \mathbb{R}^m)$ .

Using the adjoint, we can define an infinite-dimensional analogue of the Moore-Penrose pseudoinverse. Inspired by the idea that the adjoint is the infinite-dimensional analogue of the matrix transpose, we state and prove the following result.

**Theorem 4.9 (Moore-Penrose Pseudoinverse)** *Consider a bounded linear operator  $L : H_1 \rightarrow H_2$ , where  $H_1$  and  $H_2$  are Hilbert spaces over a field  $\mathbb{K}$  and  $H_2$  is finite-dimensional. Suppose  $L$  is a surjective linear operator. Then, the Moore-Penrose pseudoinverse,*

$$L^+ = L^*(LL^*)^{-1} : H_2 \rightarrow H_1, \quad (4.153)$$

*is well-defined. Further, for any  $x \in H_2$ ,  $L^+$  satisfies:*

1. *Right-inverse:*  $LL^+x = x$ .
2. *Minimum-norm:*  $\|L^+x\|_{H_1} < \|v\|_{H_1}$  for every  $v$  for which  $Lv = x$  and  $v \neq L^+x$ .

**Proof** First, we confirm that the Moore-Penrose pseudoinverse is well-defined. In particular, we'll show that  $LL^*$  is in fact invertible. First, we'll show that  $\text{range}(LL^*) = \text{range}(L)$ . One direction of this is clear—we always have  $\text{range}(LL^*) \subseteq \text{range}(L)$ . Now, let's show the reverse inclusion. In order to show  $\text{range}(LL^*) \supseteq \text{range}(L)$ , it's sufficient to show that  $\text{range}(LL^*)^\perp \subseteq \text{range}(L)$ , since the  $\perp$  “operator” reverses the direction of inclusions. Let  $z \in \text{range}(LL^*)^\perp$ . Then,  $\langle LL^*x, z \rangle = 0$  for any  $x \in H_2$ , which implies that, for  $x = z$ ,

$$0 = \langle LL^*z, z \rangle = \langle L^*z, L^*z \rangle = \|L^*z\|^2. \quad (4.154)$$

Thus, we conclude that  $z \in \ker(L^*)$ . If  $z \in \ker(L^*)$ , it must be that  $z \in \text{range}(L)^\perp$ . Thus, we conclude that  $\text{range}(LL^*)^\perp \subseteq \text{range}(L)$ . This implies that  $\text{range}(LL^*) = \text{range}(L)$ .

Now, we consider the implications of this fact for the invertibility of  $LL^*$ . Since  $LL^* : H_2 \rightarrow H_2$  is a self-adjoint map ( $(LL^*)^* = LL^*$ ) between finite-dimensional normed vector spaces, and  $\text{range}(LL^*) = \text{range}(L)$ , it follows that  $LL^*$  is positive definite ( $\langle x, LL^*x \rangle >$

$0 \forall x \neq 0 \in H_2$ ) if  $L$  is surjective. We conclude that, under the assumption of surjectivity,  $LL^*$  is invertible. The Moore-Penrose pseudoinverse is therefore well-defined.

Now, we show the two proposed items. Fix an  $x \in H_2$ . Then, we have that  $LL^+x = LL^*(LL^*)^{-1}x = x$ . Thus, the first item holds. Now, we show the second item. Suppose  $v$  satisfies  $Lv = x$  but is not equal to  $L^+x$ . To show  $\|L^+x\| < \|v\|$ , we'll first show that the difference between  $v$  and  $L^+x$  is orthogonal to  $x$ . We have,

$$\langle v - L^+x, L^+x \rangle = \langle v, L^+x \rangle - \langle L^+x, L^+x \rangle \quad (4.155)$$

$$= \langle v, L^+x \rangle - \|L^+x\|^2 \quad (4.156)$$

$$= \langle x, (LL^*)^{-1}x \rangle - \|L^+x\|^2. \quad (4.157)$$

Now, we use that  $x = LL^*(LL^*)^{-1}x$  to introduce a pseudoinverse term to the inner product. This gives,

$$= \langle LL^*(LL^*)^{-1}x, (LL^*)^{-1}x \rangle - \|L^+x\|^2 \quad (4.158)$$

$$= \langle L^*(LL^*)^{-1}x, L^*(LL^*)^{-1}x \rangle - \|L^+x\|^2 \quad (4.159)$$

$$= \|L^+x\|^2 - \|L^+x\|^2 = 0. \quad (4.160)$$

Using this result, we show that  $\|L^+x\| < \|v\|$  when  $v \neq L^+x$ . We have,

$$\|v\|^2 = \langle v, v \rangle \quad (4.161)$$

$$= \langle v - L^+x + L^+x, v - L^+x + L^+x \rangle \quad (4.162)$$

$$= \langle v - L^+x + L^+x, v - L^+x \rangle + \langle v - L^+x + L^+x, L^+x \rangle \quad (4.163)$$

$$= \|v - L^+x\|^2 + \langle L^+x, v - L^+x \rangle + \langle v - L^+x, L^+x \rangle + \|L^+x\|^2 \quad (4.164)$$

$$= \|v - L^+x\|^2 + \|L^+x\|^2, \quad (4.165)$$

where we eliminate all cross-terms using the orthogonality result from above. It follows that, when  $v \neq L^+x$ ,  $\|v\|_{H_1}^2 > \|L^+x\|_{H_1}^2$ . We conclude the desired result.  $\square$

#### 4.1.4.2 Proving the Gramian Minimum-Energy Theorem

Now, we return to the proof of Theorem 4.8. In the following proof, we rewrite the minimization problem posed in Theorem 4.8 in an operator-theoretic framework. We then use the Moore-Penrose pseudoinverse theorem we developed above to conclude the desired result.

**Proof (Of Theorem 4.8)** Fix a time  $T > 0$  and states  $x_0, x_1 \in \mathbb{R}^n$ . We'd like to show that input  $u(\cdot) \in L_2([0, T], \mathbb{R}^m)$  which minimizes the signal norm,

$$\|u\|_2^2 = \int_0^T \|u(\tau)\|_2^2 d\tau, \quad (4.166)$$

subject to the constraint  $\varphi(T, 0, x_0, u(\cdot)) = x_1$ , is given by the Gramian input,

$$u(t) = B^\top e^{A^\top(T-t)} W_c(T)^{-1} (x_1 - e^{AT} x_0). \quad (4.167)$$

Let's reframe the problem in an operator-theoretic framework. Consider the bounded linear operator,

$$L : L_2([0, T], \mathbb{R}^m) \rightarrow \mathbb{R}^n \quad (4.168)$$

$$u(\cdot) \mapsto \int_0^T e^{A(T-\tau)} B u(\tau) d\tau. \quad (4.169)$$

What we'd like to do is solve the equation,

$$x_1 = e^{AT} x_0 + Lu \iff x_1 - e^{AT} x_0 = Lu, \quad (4.170)$$

for the signal  $u \in L_2([0, T], \mathbb{R}^m)$  with minimum 2-norm. This looks exactly like the problem setup we considered above! Let's calculate the adjoint of  $L$ , and show that the Moore-Penrose pseudoinverse is well-defined under the assumption of controllability.

We recognize  $L$  as a mapping of the form considered in Example 4.3. Applying the formula we derived in this example, we conclude that

$$(L^* x)(t) = B^\top e^{A^\top(T-t)} x. \quad (4.171)$$

For such a mapping,  $LL^* : \mathbb{R}^n \rightarrow \mathbb{R}^n$  is computed,

$$LL^* x = \int_0^T e^{A(T-\tau)} B B^\top e^{A^\top(T-\tau)} x d\tau = \int_0^T e^{A\tau} B B^\top e^{A^\top \tau} d\tau x = W_c(T) x. \quad (4.172)$$

Thus,  $LL^*$  is nothing but the linear transformation corresponding to left multiplication by the controllability Gramian  $W_c(T)$ ! Since the system was assumed controllable, we know that  $W_c(T)$  must be invertible. This implies that the linear operator  $LL^*$  is invertible. We conclude that the minimum  $L_2$ -norm solution  $u(\cdot)$  to the problem,

$$x_1 - e^{AT} x_0 = Lu, \quad (4.173)$$

is given by  $u = L^*(LL^*)^{-1}(x_1 - e^{AT} x_0)$ . Using our formulas for  $L^*$  and  $LL^*$ , we conclude that the minimum  $L_2$ -norm input satisfying  $\varphi(T, 0, x_0, u(\cdot))$  is

$$u(t) = B^\top e^{A^\top(T-t)} W_c(T)^{-1} (x_1 - e^{AT} x_0). \quad (4.174)$$

This is precisely the Gramian formula we derived above. The discrete-time case follows from an identical argument in the Hilbert space  $\ell^2([0, K], \mathbb{R}^m)$  of finite-length sequences on the interval  $[0, K]$  with finite  $\ell^2$ -norm.  $\square$

**Exercise 4.11** Prove the discrete-time case of Theorem 4.8.

Let's summarize what we did in the proof above. We showed that, by re-posing our problem as a linear system in a function space, we can use the Moore-Penrose pseudoinverse to derive an optimal controller in terms of the controllability Gramian. Take a moment to appreciate all the work that went into the proof above and the (slightly miraculous) input formulas that result!

### 4.1.5 Further Reading

This section was primarily based on [36, 27, 43, 14, 6, 35]. In particular, the abstract treatment of controllability and reachability was influenced by [36] and [27], while the continuous-time controllability proofs were based on those found in [43] and [14]. The proof of the formula for the matrix exponential in terms of the first  $n - 1$  powers of  $A$  was based on that of [35]. The treatment of discrete-time controllability was primarily based on [6]. The subsection on minimum energy control closely follows the analogous section of [36].

### 4.1.6 Problems

**Problem 4.1 (Controllability of Linear, Time-Varying Systems)** In the section above, we determined sharp characterizations of controllability for linear, time-invariant system representations. Determine a necessary and sufficient Gramian-like condition for a continuous-time, LTV representation  $(A(\cdot), B(\cdot), C(\cdot), D(\cdot))$  to be controllable.

**Problem 4.2 (Block Controllability [35])** Consider the  $n$ -dimensional linear, time-invariant system with state equation,

$$\dot{x}(t) = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} x(t) + \begin{bmatrix} B_{11} \\ 0 \end{bmatrix} u(t), \quad (4.175)$$

where  $A_{11} \in \mathbb{R}^{q \times q}$  and  $B_{11} \in \mathbb{R}^{q \times m}$  with  $\text{rank } q$ . Prove that this system is controllable if and only if the  $(n - q)$ -dimensional linear system,

$$\dot{x}(t) = A_{22}z(t) + A_{21}v(t), \quad (4.176)$$

is controllable.

**Problem 4.3 (Piecewise Constant Control [3])** The *double integrator* is the system with state equation,

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u. \quad (4.177)$$

Find a piecewise constant control strategy that drives the system from the origin to the state  $(1, 1)$  at a time  $T$ .

**Problem 4.4 (A Diagonal Controllability Condition [36])** Consider the linear, time-invariant system with state equation,

$$\dot{x}(t) = \begin{bmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_n \end{bmatrix} x(t) + \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} u(t), \quad (4.178)$$

where  $x(t) \in \mathbb{R}^n$  and  $u(t) \in \mathbb{R}$ . Show that the system is controllable if and only if  $\lambda_i \neq \lambda_j$ ,  $i \neq j$ , and  $b_i \neq 0$ .

**Problem 4.5 (Controllability Miscellanea [9, 25])** Consider a linear, time-invariant system with state equation  $\dot{x}(t) = Ax(t) + Bu(t)$ ,  $x(t) \in \mathbb{R}^n$ ,  $u(t) \in \mathbb{R}^m$ . For each of the following statements, provide either a proof or a counterexample.

1. Suppose the pair  $(A, B)$  is controllable. Is the pair  $(A^2, B)$  controllable?
2. Suppose the pair  $(A^2, B)$  is controllable. Is the pair  $(A, B)$  controllable?
3. Suppose  $(A, B)$  is controllable. For a nonzero initial condition  $x(0) = x_0 \neq 0$ , is it possible to find a piecewise continuous input  $u(\cdot) \in PC(\mathbb{R}_{\geq 0}, \mathbb{R}^m)$  such that the system is *brought to rest* at  $t = 1$  (i.e.  $x(t) = 0$  for all  $t \geq 1$ )?
4. Suppose  $(A, B)$  is controllable. Fix a state  $\bar{x} \in \mathbb{R}^n$ . If the system is initially at rest ( $x(0) = 0$ ), does there exist a piecewise continuous input  $u(\cdot) \in PC(\mathbb{R}_{\geq 0}, \mathbb{R}^m)$  for which  $x(t) = \bar{x}$  for all  $t \geq 1$ ?

**Problem 4.6 (Minimum Energy Requires to Leave Safe Operating Region [6])**

Consider the linear, time-invariant system  $\dot{x} = Ax + Bu$ ,  $x(0) = 0$ , where  $x(t) \in \mathbb{R}^n$ ,  $u(t) \in \mathbb{R}^m$ . Suppose  $A$  is Hurwitz and that  $(A, B)$  is a controllable pair. Consider the following scenario. Suppose we *don't* have any control over the value of the input signal  $u(\cdot)$ , but we have some idea of how large its total energy,

$$\|u\|_2^2 = \int_0^\infty \|u(\tau)\|_2^2 d\tau, \quad (4.179)$$

is likely to be. The *safe operating region* for the system is the ball,

$$\mathcal{B} = \{x \in \mathbb{R}^n : \|x\|_2 \leq 1\}. \quad (4.180)$$

The hope is that the (unknown) input signal  $u(\cdot)$  will not drive the state of the system outside the safe operating region. One measure of system security that is used is the minimum energy  $E_{\min}$  that is required to drive the state outside the safe operating region. Fixing  $x_0 = 0$  for simplicity, we define,

$$E_{\min} = \inf_{t \in \mathbb{R}_{>0}, u(\cdot) \in \mathcal{U}} \int_0^t \|u(\tau)\|_2^2 d\tau \text{ s.t. } \varphi(t, 0, 0, u(\cdot)) \notin \mathcal{B}. \quad (4.181)$$

Notice that we *do not* fix the time  $t$  at which the state leaves the safe operating region—rather,  $t$  is an optimization variable. If  $E_{\min}$  is much larger than the energy of the unknown input signals we can expect, we can be fairly confident that the state will not leave the safe operating region.

1. Fix an  $x \in \mathbb{R}^n$ . Recall that, for each  $t \in \mathbb{R}_{>0}$ , the solution of the optimization problem,

$$\arg \min_{u(\cdot) \in \mathcal{U}} \int_0^t \|u(\tau)\|_2^2 d\tau, \text{ s.t. } x(0) = 0, x(t) = x, \quad (4.182)$$

is given by the input signal  $u(\tau) = B^\top e^{A^\top(t-\tau)} W_c(t)^{-1} x$ , where  $W_c(t)$  is the controllability Gramian at time  $t$ . Using this fact, show that the infimum,

$$\inf_{t \in \mathbb{R}_{>0}, u(\cdot) \in \mathcal{U}} \int_0^t \|u(\tau)\|_2^2 d\tau, \text{ s.t. } x(0) = 0, \varphi(t, 0, 0, u(\cdot)) = x, \quad (4.183)$$



is equal to  $x^\top W_c^{-1}x$ , where  $W_c = \lim_{t \rightarrow \infty} W_c(t)$  is the infinite-horizon controllability Gramian.

2. Using your answer to part (1), calculate  $E_{\min}$ . Your solution should be in terms of the matrices  $A, B$ , or other matrices derived from them such as the controllability matrix  $\mathcal{C}$ , the infinite-horizon controllability Gramian  $W_c$ , and its inverse  $P = W_c^{-1}$ . Simplify your answer as completely as you can.
3. Suppose the safe operating region is the unit cube  $C = \{x \in \mathbb{R}^n : |x_i| \leq 1, i = 1, \dots, n\}$  instead of the unit ball  $\mathcal{B}$ . Let  $E_{\min}^{\text{cube}}$  denote the minimum energy required to drive the state outside the unit cube  $C$ . Repeat part (2) for  $E_{\min}^{\text{cube}}$ . Once again, simplify your answer as completely as you can.



## References

1. Stephen Abbott. *Understanding analysis*, volume 2. Springer, 2001.
2. Panos J Antsaklis and Anthony N Michel. *Linear systems*, volume 8. Springer, 1997.
3. Karl Johan Åström and Richard Murray. *Feedback systems: an introduction for scientists and engineers*. Princeton university press, 2021.
4. Sheldon Axler. *Measure, integration & real analysis*. Springer Nature, 2020.
5. Sheldon Axler. *Linear algebra done right*. Springer Nature, 2024.
6. Stephen Boyd. *Lecture Notes for EE263*. Stanford University, Available online at [https://ee263.stanford.edu/archive/ee263\\_course\\_reader.pdf](https://ee263.stanford.edu/archive/ee263_course_reader.pdf), 2008.
7. Stephen Boyd and CA Desoer. Subharmonic functions and performance bounds on linear time-invariant feedback systems. *IMA Journal of Mathematical control and Information*, 2:153–170, 1985.
8. Frank M Callier and Charles A Desoer. *Linear system theory*. Springer Science & Business Media, 2012.
9. Chih-Yuan Chiu, Claire Tomlin, and Yi Ma. *Linear Systems*. Available online at [https://ucb-ee106.github.io/106b-sp23site/assets/Linear\\_Systems\\_Professor\\_Ma.pdf](https://ucb-ee106.github.io/106b-sp23site/assets/Linear_Systems_Professor_Ma.pdf), 2019.
10. Mohammed Dahleh, Munther A Dahleh, and George Verghese. *Lectures on dynamic systems and control*. Massachusetts Institute of Technology, 2004.
11. John C. Doyle. Analysis of feedback systems with structured uncertainties. In *IEEE Proceedings D (Control Theory and Applications)*, volume 129, pages 242–250. IET Digital Library, 1982.
12. John C. Doyle, Bruce A Francis, and Allen R Tannenbaum. *Feedback control theory*. Courier Corporation, 2013.
13. John C. Doyle, Keith Glover, Pramod Khargonekar, and Bruce Francis. State-space solutions to standard  $h_2$  and  $h_\infty$  control problems. In *1988 American Control Conference*, pages 1691–1696. IEEE, 1988.
14. Geir E Dullerud and Fernando Paganini. *A course in robust control theory: a convex approach*, volume 36. Springer Science & Business Media, 2013.
15. Stephen H Friedberg, Arnold J Insel, and Lawrence E Spence. *Linear Algebra*. Pearson, 2014.
16. Michael Green and David J.N. Limebeer. *Linear Robust Control*. Dover, 1995.
17. Morris W Hirsch, Stephen Smale, and Robert L Devaney. *Differential equations, dynamical systems, and an introduction to chaos*. Academic press, 2013.
18. Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Dover, 1985.
19. Haimin Hu. *Computing Jordan Form Using Jordan Chain*. Online Distribution, UC Berkeley EE221A, 2017.
20. Hassan K. Khalil. *Nonlinear Systems*. Prentice Hall, 2002.
21. Edward A. Lee and Pravin Varaiya. *Structure and Interpretation of Signals and Systems, Second Edition*. Available online at <https://ptolemy.berkeley.edu/books/leearaiya/>, 2011.
22. Andrew D. Lewis. *A Mathematical Approach to Classical Control*. Online Distribution, 2016.
23. Andrew D. Lewis. *Introduction to Differential Equations*. Online Distribution, 2017.
24. Andrew D. Lewis. *A Mathematical Introduction to Signals and Systems, Volume 4*. Available online at <https://mast.queensu.ca/~andrew/teaching/SigSys/pdf/volume4.pdf>, 2022.

25. John Lygeros and Federico Ramponi. Lecture notes on linear system theory. *Automatic Control Laboratory, ETH Zurich*, 2010.
26. Alexandre Megretski. *Multivariable Control Systems*. Available online at <https://web.mit.edu/6.245/www/schedule.html>, 2011.
27. Richard M Murray. *Feedback Systems: Notes on Linear Systems Theory*. Available online at [https://fbswiki.org/wiki/index.php/Supplement:\\_Linear\\_Systems\\_Theory](https://fbswiki.org/wiki/index.php/Supplement:_Linear_Systems_Theory), 2020.
28. Alan V. Oppenheim and Alan S. Willsky. *Signals and Systems*. Prentice Hall, 1997.
29. Andrew Packard and John C. Doyle. The complex structured singular value. *Automatica*, 29(1):71–109, 1993.
30. Andrew Packard, Roberto Horowitz, Kameshwar Poola, and Francesco Borrelli. *Dynamic Systems and Feedback, Class Notes*. Available online, 2018.
31. Lawrence Perko. *Differential Equations and Dynamical Systems*. Springer, 2000.
32. Ian Postlethwaite and Sigurd Skogestad. *Multivariable Feedback Control, Analysis & Design*. Wiley, 2005.
33. Walter Rudin. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1964.
34. Walter Rudin. *Real and Complex Analysis*. McGraw-Hill, 1987.
35. Wilson J. Rugh. *Linear System Theory*. Prentice Hall, 1996.
36. Eduardo D Sontag. *Mathematical control theory: deterministic finite dimensional systems*, volume 6. Springer Science & Business Media, 2013.
37. Elias M Stein and Rami Shakarchi. *Real analysis: measure theory, integration, and Hilbert spaces*, volume 3. Princeton University Press, 2009.
38. Elias M Stein and Rami Shakarchi. *Complex analysis*, volume 2. Princeton University Press, 2010.
39. Elias M Stein and Rami Shakarchi. *Fourier analysis: an introduction*, volume 1. Princeton University Press, 2011.
40. Masayoshi Tomizuka. *Advanced Control Systems I*. Online Distribution, 2022.
41. Fawwaz T. Ulaby and Andrew E. Yagle. *Signals and Systems: Theory and Applications*. Available online at <https://ss2-2e.eecs.umich.edu/>, 2018.
42. Kemin Zhou and John C. Doyle. *Essentials of robust control*. Prentice hall, 1998.
43. Kemin Zhou, John C. Doyle, and Keith Glover. *Robust and Optimal Control*. Prentice hall, 1996.