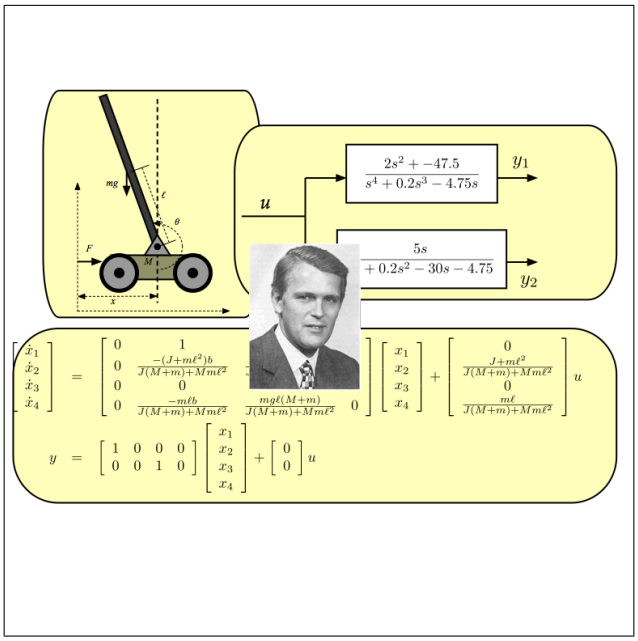
# Lectures on Linear Systems Theory

**December 2, 2024**

Department of Electrical Engineering
University of Notre Dame



# Contents

# Preface

This book grew out of lectures I gave for a semester-long course in *linear systems theory* for first year engineering graduate students at the University of Notre Dame. This is a course in applied mathematics. Most students may only need their prior undergraduate work in transform methods, linear algebra Strang (1976) , and differential equations. The lectures, however, should also be of interest to those students with a more mature mathematical sensibility. The course is structured similar to classical textbooks in linear systems theory such as Kailath (1980) and Antsaklis and Michel (2006), as well as the more recent textbook, Hespanha (2018). The lecture notes make extensive use of MATLAB and the control systems toolbox. These notes are a work in progress, having been revised and reorganized several times over the past decade.

M. D. Lemmon
Department of Electrical Engineering
University of Notre Dame
Summer, 2024

CHAPTER 1

# Mathematics for Linear Systems

**Linear systems theory** applies linear algebra concepts to the study of dynamical systems that can be modeled by linear differential equations. In this regard, linear systems theory is a branch of applied mathematics. Linear systems theory, however, also provides a set of tools that engineers commonly use to predict how a physical process might behave. Many undergraduate level engineers are familiar with the use of transform-based methods in solving systems of linear differential equations. These undergraduates are also familiar with the use of transfer functions to predict how a circuit or mechanical system might respond to a known input signal. These methods are specific tools showing how linear systems theory is used in engineering. The purpose of this course is to take a deeper look at how linear system theory is used in the modeling of dynamical systems found in many engineering disciplines.

This chapter reviews the mathematical concepts needed to take that deep dive into linear systems theory. In particular, we start by examining conditions for the existence and uniqueness of solutions to linear algebraic equations. These conditions can be formulated in terms of linear algebra concepts. So we then review basic linear algebra (linear space, linear transformations, eigendecompositions) and look at other useful results from linear algebra (singular value decompositions and the Cayley-Hamilton theorem [Strang (1976)]).

## 1. Linear Algebraic Equations

**A system of linear algebraic equations (LAE)** is a matrix-vector equation of the form

$$b = \mathbf{A}x \tag{1}$$

where $x \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, and $\mathbf{A} \in \mathbb{R}^{m \times n}$. The problem is to find the vector $x$ such that $b = \mathbf{A}x$, assuming we already know the vector $b$ and the matrix $\mathbf{A}$. Any such vector, $x$, is called a *solution* of the LAE. A number of problems in the theory of linear dynamical systems can be reduced to solving a system of LAEs. We will therefore find it useful to ask the following three questions about equation (1).

(1) *Existence:* Does a solution exist?
(2) *Uniqueness:* Is there more than one solution to the LAE?
(3) *Computatability:* How does one compute a solution to the LAE?

This lecture answers these questions with respect to concrete examples.

### 1.1. Gaussian Elimination With Back Substitution:

Let us start with a numerical procedure that is often used to *compute* a solution to an LAE. This procedure is called *Gaussian Elimination with Back Substitution*. Consider the following system of linear algebraic equations

$$\begin{array}{rcrcrcr} 1 & = & 2x_1 & + & x_2 & + & x_3 \\ -2 & = & 4x_1 & + & x_2 & & \\ 7 & = & -2x_1 & + & 2x_2 & + & x_3 \end{array} \tag{2}$$

It will be convenient to rewrite the preceding equations as a single matrix-vector equation of the form,

$$b = \begin{bmatrix} 1 \\ -2 \\ 7 \end{bmatrix} = \begin{bmatrix} 2 & 1 & 1 \\ 4 & 1 & 0 \\ -2 & 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \mathbf{A}x$$

where $\mathbf{A} \in \mathbb{R}^{3\times 3}$ is a matrix whose elements are the coefficients in equation (2) and $x \in \mathbb{R}^3$ is the vector whose $i$th element , $x_i$ ($i = 1, 2, 3$), is the $i$th element of the triple solving the system of equations. Our problem is to find a solution to this system of equations.

The strategy used to compute a solution is a recursive process that consists of two parts. The first part is called the *elimination* phase and it systematically eliminates variables from a subset of equations to identify a subproblem of lower dimension. The second part is called *back substitution* and it systematically uses any solution for the lower dimensional subproblem to find a solution for the eliminated variables in the higher order problem. The elimination strategy is applied in a recursive manner to an $n$-dimensional system. This strategy reduces the $n$-dimensional problem to an $n-1$ dimensional problem, and continues to reduce the problem size until one has a 1-dimensional subproblem. The 1-dimensional subproblem is trivial to solve so we use this to "bootstrap" up to the $n$-dimensional solution through back substitution. In particular, back substitution is used on the $k-1$ dimensional problem to obtain the solution to the $k$ dimensional problem where $k = 2, 3, \ldots, n$. This recursive strategy is what we refer to as *variable elimination with back substutition*. The particular elimination strategy we will use is called *Gaussian elimination*.

We will use the system of equation (2) to illustrate the Gaussian elimination phase of the method. One first reduces the problem from a system of equations with 3 unknowns (3-dimensional problem) to a smaller 2-dimensional problem. Gaussian elimination uses the sequential application of *elementary row operations* to achieve this reduction. An elementary row operation is a transformation on the problem equations in which

- The order of two equations in the entire problem is reversed.
- One equation is multiplied by a real number and the result replaces the originally selected equation.

- One equation is added to another equation and that second equation is replaced by the sum.

These elementary row operations are invoked in a systematic manner that takes a set of $k$ equations ($k = 2, \ldots, n$), eliminates a given variable from $k-1$ of equations to obtain a smaller set of equations to solve.

For example, let us consider the 3-dimensional system in equation (2). Let us multiply the first equation by $-2$, add the resulting equation to the second equation, and then replace the second equation with that sum. This operation removes the variable $x_1$ from the second equation, thereby only making it a function of $x_2$ and $x_3$. The resulting system of equations is

$$\begin{array}{rcrcrcr} 1 & = & 2x_1 & + & x_2 & + & x_3 \\ -4 & = & & - & x_2 & - & 2x_3 \\ 7 & = & -2x_1 & + & 2x_2 & + & x_3 \end{array} \tag{3}$$

We still have $x_1$ in the third equation. So let us use row operations to remove equation 3's dependence on $x_1$. In particular, this can be done by adding the first equation to the third equation and replace the third equation with the result. This sequence of row operations gives

$$\begin{array}{rcrcrcr} 1 & = & 2x_1 & + & x_2 & + & x_3 \\ -4 & = & & - & x_2 & - & 2x_3 \\ 8 & = & & & 3x_2 & + & 2x_3 \end{array} \tag{4}$$

One may readily verify that the variable $x_1$ does not appear in the last two equations. So, if we already knew that $x_2$ and $x_3$ satisfy the last two equations

$$\begin{array}{rcrcr} -4 & = & -x_2 & - & 2x_3 \\ 8 & = & 3x_2 & + & 3x_3 \end{array} \tag{5}$$

Then one could use the first equation to rewrite $x_1$ as a function of $x_2$ and $x_3$. In particular, simple algebra shows that

$$x_1 = \frac{1}{2}(1 - x_2 - x_3) \tag{6}$$

thereby giving us the solution of the full LAE provided we already know what $x_2$ and $x_3$ are. The two-step reduction is the first stage of this Gaussian elimination process. The update in equation (6) represents the back substitution phase of the algorithm.

Note that the preceding procedure reduced the problem of finding the solution of the 3-dimensional system to that of finding the solution of the associated 2-dimensional system in equation (5). So let us apply this procedure one more time to reduce the 2-dimensional system in equation (5) to a 1-dimensional system. Note that the one dimensional system is trivial to solve. This last elimination is designed to remove $x_2$ from the third equation. In particular, if we multiply the last equation in equation (5) by 3, add it to the second equation in (5), and replacing gives

$$\begin{array}{rcrcr} -4 & = & -x_2 & - & 2x_3 \\ -4 & = & & - & 4x_3 \end{array} \tag{7}$$

The last equation is

$$-4 = -4x_3$$

which is a 1-dimensional system of equations whose solution is readily seen to be $x_3 = 1$. We back substitute this value for $x_3$ in the first line of equation (7) to obtain

$$-4 = -x_2 - 2$$

whose solution is readily seen to be $x_2 = 2$. We now know that $x_2 = 2$ and $x_3 = 1$, so we back substitute these values into equation (6) to obtain

$$x_1 = \frac{1}{2}(1 - x_2 - x_3) = \frac{1}{2}(1 - 2 - 1) = -1$$

which completes the computed solution as $x = \begin{bmatrix} -1 \\ 2 \\ 1 \end{bmatrix}$, whose correctness is readily checked by multiplying out $\mathbf{A}x$ and verifying that it is equal to the given $b$ vector.

**Remark:** The application of the row operations described above transforms the original system of equations (2) to the following form

$$\begin{array}{rcrcrcr} 1 & = & 2x_1 & + & x_2 & + & x_3 \\ -4 & = & & - & x_2 & - & 2x_3 \\ -4 & = & & & & - & 4x_3 \end{array} \tag{8}$$

which can be written in matrix vector form as

$$\begin{bmatrix} 1 \\ -4 \\ -4 \end{bmatrix} = \begin{bmatrix} 2 & 1 & 1 \\ 0 & -1 & -2 \\ 0 & 0 & -4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \tag{9}$$

Note that the matrix in this equation is in *upper triangular form* where all elements below the diagonal are zero. Gaussian elimination may therefore be seen as a *transforming* the original system of equations to an upper triangular form, from which back substitution is applied. In particular, each of the row operations given above can be realized as a matrix-vector multiplication on the original system,

$$\mathbf{P}b = \mathbf{P}\mathbf{A}x$$

where

$$\begin{aligned} \mathbf{P} &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 3 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -5 & 3 & 1 \end{bmatrix} \end{aligned}$$

Each matrix in the first line corresponds to a particular row operation described above.

Does the transformed system of equations have the same solution as the original system? To answer this question let $x$ solve

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

Let $\mathbf{P}$ be any nonsingular matrix and let $z$ solve the system

$$\mathbf{P}b = \mathbf{P}\mathbf{A}z$$

Since $\mathbf{P}$ is invertible, we know there exists a matrix inverse, $\mathbf{P}^{-1}$. Multiply the above equation by $\mathbf{P}^{-1}$ to obtain

$$\mathbf{P}^{-1}\mathbf{P}b = \mathbf{P}^{-1}\mathbf{P}\mathbf{A}z$$

Since $\mathbf{P}^{-1}\mathbf{P} = \mathbf{I}$ (the identity matrix) we can readily see the above equation reduces to

$$b = \mathbf{A}z$$

which means the solution to the transformed $z$-system also is a solution for the original $x$-system. We can therefore conclude that applying any nonsingular transformation to the system will not change its "solution".

**1.2. Existence and Uniqueness of Solutions.** Let us now examine when a *unique* solution exists for a system of linear algebraic equations. We can do this by seeing when the Gaussian elimination procedure begins to break down. In particular, consider the following system that has been written in matrix-vector form,

$$b = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 & 3 & 3 & 2 \\ 2 & 6 & 9 & 5 \\ -1 & -3 & 3 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \mathbf{A}x \tag{10}$$

There are two interesting things about this system. First notice that $b = \overline{0}$. Such systems are said to be *homogeneous*. The second interesting thing is that there are more unknown variables than equations. When this occurs, the system is said to be *under-determined*

Let us apply the Gaussian elimination procedure and see what happens. Applying row operations to null the $(2,1)$ and $(3,1)$ components of the $\mathbf{A}$

matrix gives

$$\mathbf{A} = \begin{bmatrix} 1 & 3 & 3 & 2 \\ 2 & 6 & 9 & 5 \\ -1 & -3 & 3 & 0 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 3 & 3 & 2 \\ 0 & 0 & 3 & 1 \\ -1 & -3 & 3 & 0 \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} 1 & 3 & 3 & 2 \\ 0 & 0 & 3 & 1 \\ 0 & 0 & 6 & 2 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 3 & 3 & 2 \\ 0 & 0 & 3 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} = \mathbf{U}$$

Examining the first row of matrix $\mathbf{U}$, we see that the pivot occurs in the row column so the associated variable is $x_1$. As discussed above, this means one can express $x_1$ in terms of $x_2$, $x_3$, and $x_4$. Examining the second row of $\mathbf{U}$, we see a pivot of 3 in the third column. So the associated variable is $x_3$ and this means that we can express $x_3$ in terms of $x_4$.

The last row of $\mathbf{U}$ is zero, which means there are no nonzero pivots associated with this row. In particular, this means that the only variables with nonzero pivots are $x_1$ and $x_3$. The other two variables, $x_2$ and $x_4$, have no nonzero pivots. When this occurs we say the system of linear algebraic equations is *singular*. When a system is singular then it either has no solution or an infinite number of solutions.

On the basis of the preceding discussion, one groups the variables in $x$ into two disjoint sets. The first set consists of *basic variables* that correspond to variables with nonzero pivots. In the above example, the basic variables are $x_1$ and $x_3$. Variables that are not basic are said to be *free variables*. In our example, the free variables are $x_2$ and $x_4$. Free variables cannot be expressed in terms of the other variables of the equation.

To find the most general solution, one allows the free variables to take any value and then uses back substitution to express the basic variables in terms of the free variables. For this example, the upper triangular system of

equations after the elimination procedure is

$$\begin{array}{rcccccccc} 0 & = & x_1 & + & 3x_2 & + & 3x_3 & + & 2x_4 \\ 0 & = & & & & & 3x_3 & + & x_4 \end{array}$$

In the second equation, the basic variable is $x_3$ that we rewrite as a function of the free variable, $x_4$,

$$x_3 = -\frac{1}{3}x_4$$

Inserting this algebraic expression for $x_3$ into the first equation and solving for the remaining basic variable, $x_1$, in terms of the free variables gives

$$x_1 \;\; = \;\; -3x_2 - x_4$$

So all solutions to this particular system of equations may be written as

$$x = \begin{bmatrix} -3x_2 - x_4 \\ x_2 \\ -\frac{1}{3}x_4 \\ x_4 \end{bmatrix} = x_2 \begin{bmatrix} -3 \\ 1 \\ 0 \\ 0 \end{bmatrix} + x_4 \begin{bmatrix} -1 \\ 0 \\ -\frac{1}{3} \\ 1 \end{bmatrix}$$

where $x_2$ and $x_4$ are free to be any real numbers. All solutions for this under-determined homogeneous system may therefore be expressed as a *linear combination* of two vectors

$$\left\{ \begin{bmatrix} -3 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \\ -\frac{1}{3} \\ 1 \end{bmatrix} \right\}.$$

This set of solutions is said to form a *subspace* of the Euclidean vector space $\mathbb{R}^4$ and the two vectors form a *basis* set for that subspace. One may also view the linear system of equations as a *linear transformation* mapping vectors in $\mathbb{R}^4$ onto vectors in $\mathbb{R}^3$. In particular, the matrix $\mathbf{A}$ in the homogeneous equation $\mathbf{A}x = 0$ maps $x \in \mathbb{R}^4$ onto the zero vector in $\mathbb{R}^3$. The set of vectors in $\mathbb{R}^4$ that are mapped onto the zero vector through $\mathbf{A}$, form another useful subspace called the *null space* of the matrix $\mathbf{A}$. The null space is sometimes called the *kernel* of the matrix and is often denoted as $\ker(\mathbf{A})$. In particular, this means that any solution to the homogeneous

equation $\mathbf{A}x = 0$ must lie in the null space of $\mathbf{A}$. We can therefore say that all solutions of the homogeneous equation are $x \in \ker(\mathbf{A})$.

The preceding discussion characterized the solutions when $b = \overline{0}$. If $b$ is nonzero then we have an *inhomogeneous* system of equations. In particular, if one wants to find solutions to the inhomogeneous problem, we apply Gaussian elimination to the augmented matrix $\left[\begin{array}{c|c} \mathbf{A} & b \end{array}\right]$. So let us consider the inhomogeneous version of our above example where we let the elements of $b$ be real variables. So the system of equations can be written as

$$b = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 1 & 3 & 3 & 2 \\ 2 & 6 & 9 & 5 \\ -1 & -3 & 3 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \mathbf{A}x$$

The matrix tableaux for this system and its subsequent reduction to reduced row echelon form generate are shown below

$$\left[\begin{array}{cccc|c} 1 & 3 & 3 & 2 & b_1 \\ 2 & 6 & 9 & 5 & b_2 \\ -1 & -3 & 3 & 0 & b_3 \end{array}\right] \Rightarrow \left[\begin{array}{cccc|c} 1 & 3 & 3 & 2 & b_1 \\ 0 & 0 & 3 & 1 & -2b_1 + b_2 \\ 0 & 0 & 6 & 2 & b_1 + b_3 \end{array}\right]$$

$$\Rightarrow \left[\begin{array}{cccc|c} 1 & 3 & 3 & 2 & b_1 \\ 0 & 0 & 3 & 1 & b_2 - 2b_1 \\ 0 & 0 & 0 & 0 & b_3 - 2b_2 + 5b_1 \end{array}\right]$$

Note that the last equation requires

$$0 \;=\; 5b_1 - 2b_2 + b_3$$

This equation is satisfied when $b = 0$ and when $b = \begin{bmatrix} 1 \\ 1 \\ -3 \end{bmatrix}$. But it will not be satisfied for any arbitrary choice for $b$. For instance, the equation is not satisfied if $b = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$. This would mean, therefore, that a solution to

the inhomogeneous problem *does not exist* when $b = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$. So we are interested in determining all of those vectors $b \in \mathbb{R}^3$ for which a solution does exist.

To determine which $b$'s in $\mathbb{R}^3$ give rise to an inhomogeneous system with real solutions, we first note that $b$ can be written as a linear combination of the *columns* of $\mathbf{A}$,

$$\begin{aligned} b &= \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 1 & 3 & 3 & 2 \\ 2 & 6 & 9 & 5 \\ -1 & -3 & 3 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \mathbf{A}x \\ &= x_1 \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix} + x_2 \begin{bmatrix} 3 \\ 6 \\ -3 \end{bmatrix} + x_3 \begin{bmatrix} 3 \\ 9 \\ 3 \end{bmatrix} + x_4 \begin{bmatrix} 2 \\ 5 \\ 0 \end{bmatrix} \end{aligned} \tag{11}$$

These $b$ vectors, therefore, lie in a subspace of $\mathbb{R}^3$ that is *spanned* by the *columns* of the $\mathbf{A}$ matrix. We call this subspace the *column space* of $\mathbf{A}$. If we think of $\mathbf{A}$ as a linear transformation, then it will also be called the *range space* of $\mathbf{A}$. We often denote this range space as $\mathrm{range}(\mathbf{A})$. Our preceding discussion has therefore shown that a solution *exists* for the inhomogeneous system of linear equations $b = \mathbf{A}x$ if and only if

$$b \in \mathrm{range}(\mathbf{A}) = \text{range space of } \mathbf{A}$$

By inspection of the four vectors used to form $b$ in equation (11), one can see that only two of these vectors are *linearly independent* (i.e. they cannot be written as a linear combination of the other vectors). This means that

$$\mathrm{range}(\mathbf{A}) = \mathrm{span} \left\{ \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix}, \begin{bmatrix} 1 \\ 3 \\ 1 \end{bmatrix} \right\} \tag{12}$$

where $\mathrm{span}\{z_1, \ldots, z_n\}$ is the subspace formed by all linear combinations of the collection of vectors $z_1$ to $z_n$. Because all elements of the range space defined by equation (12) are linearly independent, we know this is the smallest number of vectors that can be used to span $\mathrm{range}(\mathbf{A})$. We refer to such a collection of vectors as a *basis set* and the number of elements in the basis is called the *dimension* of the subspace $\mathrm{range}(\mathbf{A})$. We also refer to this as the *rank* of the matrix $\mathbf{A}$.

Clearly not all $b \in \mathbb{R}^3$ will lie in $\mathrm{range}(\mathbf{A})$. But for those that do, one may solve the inhomogeneous problem using back substitution. For this particular example, one can readily verify that all solutions are

$$x = x_2 \begin{bmatrix} -3 \\ 1 \\ 0 \\ 0 \end{bmatrix} + x_4 \begin{bmatrix} -1 \\ 0 \\ -\frac{1}{3} \\ 1 \end{bmatrix} + \begin{bmatrix} 3b_1 - b_2 \\ 0 \\ \frac{1}{3}(-2b_1 + b_2) \\ 0 \end{bmatrix}$$

The first two terms on the right hand side of the above equation are vectors forming a basis for the null space of $\mathbf{A}$. The third term on the right hand side is a *particular solution* of the original inhomogeneous problem. In particular, this means that if $b \in \mathrm{range}(\mathbf{A})$ then any solution to this problem may be written as

$$x \in x_p + \ker(\mathbf{A})$$

where $x_p$ is a particular solution to the system and $\ker(\mathbf{A})$ is the null space of $\mathbf{A}$.

We can now answer the questions we originally posed.

(1) **Existence?** A solution exists if $b$ lies in the range space of $\mathbf{A}$. In other words, $b \in \mathrm{range}(\mathbf{A})$.
(2) **Uniqueness ?** If $b \in \mathrm{range}(\mathbf{A})$, then any solution can be written as $x_p + v$ where $x_p$ is any particular solution such that $\mathbf{A}x_p = b$ and $v$ is any vector in the null space of $\mathbf{A}$. The solution, therefore, will

be unique if and only if $\ker(\mathbf{A})$ is the trivial linear space consisting of only the zero vector.

(3) **Computability?** The Gaussian elimination procedure with back substitution provides an efficient algorithm for computing solutions to the inhomogeneous problem.

**1.3. Relaxed Solution Concepts.** There are many real-life engineering problems giving rise to a system of linear equations $b = \mathbf{A}x$ where $b \notin \text{range}(\mathbf{A})$. Based on our earlier discussion, this means that no solution exists for this system of equations. But this fact seems to contradict the fact that we "know" our real-life problem does have a solution. This commonly occurs in LAEs that have more equations than variables. Such LAEs are said to be *over-determined* and they often occur in parameter estimation problems where one relies on noisy measurements to "estimate" a vector of unknown parameters. Because there are many more equations than unknowns, and because of the noise, it is highly unlikely that $b$ lies in the range space of $\mathbf{A}$. So what can we do?

Let us consider the following over-determined LAE,

$$b = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \mathbf{A}x$$

As mentioned above, it is highly unlikely that $b$ will be in the range space of $\mathbf{A}$, so let us *relax* our notion of a solution so that $x$ solves the LAE if $\mathbf{A}\widehat{x} - b \approx 0$. In other words, we don't require that $x$ is an exact solution in the sense that $\mathbf{A}x = b$, but only that $\mathbf{A}x$ is "close" to $b$. We need to clarify what it means to be "close". Formally, we define this as requiring that

$$|\mathbf{A}x - b|^2 \leq \epsilon$$

where $|x| = \sqrt{x^T x}$ is the Euclidean norm of vector $x$ and $\epsilon$ is a "small" tolerance level that quantifies how close $\mathbf{A}x$ is to $b$. In particular we want

to find a solution $\widehat{x}$ that minimizes $\epsilon$. We can pose our search for this minimizing $\widehat{x}$ as an optimization problem

$$\widehat{x} \;=\; \arg\min_{x\in\mathbb{R}^n} |\mathbf{A}x-b|^2$$

We can find a solution by writing out an explicit expression for

$$\begin{aligned} J(x) \;&\stackrel{\text{eq}}{=}\; \frac{1}{2}|\mathbf{A}x-b|^2 = \frac{1}{2}\sum_{i=1}^{3}\left(b_i-\overline{a}_i^T x\right)^2 \\ &=\; \frac{1}{2}\sum_{i=1}^{3}\left(b_i-\left[\begin{array}{cc} a_{i1} & a_{i2}\end{array}\right]\left[\begin{array}{c} x_1\\ x_2\end{array}\right]\right)^2 \end{aligned}$$

where $\overline{a}_i$ is the $i$th row of $\mathbf{A}$. We can find the $x$ that minimizes the above expression by taking the derivative of the right hand side of the above equation and setting it equal to zero. From elementary calculus, we know this is a necessary condition for optimality. Taking the derivatives means the optimal $\widehat{x}$ satisfies

$$\begin{aligned} 0 \;&=\; \frac{\partial J}{\partial \widehat{x}_1} = -\sum_{i=1}^{3}\left(b_i-\left[\begin{array}{cc} a_{i1} & a_{i2}\end{array}\right]\left[\begin{array}{c} \widehat{x}_1\\ \widehat{x}_2\end{array}\right]\right)a_{i1} \\ 0 \;&=\; \frac{\partial J}{\partial \widehat{x}_2} = \sum_{i=1}^{3}\left(b_i-\left[\begin{array}{cc} a_{i1} & a_{i2}\end{array}\right]\left[\begin{array}{c} \widehat{x}_1\\ \widehat{x}_2\end{array}\right]\right)a_{i2} \end{aligned}$$

which we can rewrite in matrix vector form as

$$\begin{aligned} 0 \;&=\; -\left[\begin{array}{ccc} a_{11} & a_{21} & a_{31}\\ a_{12} & a_{22} & a_{32}\end{array}\right]\left[\begin{array}{c} b_1\\ b_2\\ b_3\end{array}\right]+\left[\begin{array}{ccc} a_{11} & a_{21} & a_{31}\\ a_{12} & a_{22} & a_{32}\end{array}\right]\left[\begin{array}{cc} a_{11} & a_{12}\\ a_{21} & a_{22}\\ a_{31} & a_{32}\end{array}\right]\left[\begin{array}{c} \widehat{x}_1\\ \widehat{x}_2\end{array}\right] \\ &=\; -\mathbf{A}^T b+\mathbf{A}^T\mathbf{A}\widehat{x} \end{aligned}$$

Note that if $\mathbf{A}^T\mathbf{A}$ is invertible then the solution to this problem is

$$\widehat{x} = \left[\mathbf{A}^T\mathbf{A}\right]^{-1}\mathbf{A}^T b$$

with the minimum value being

$$\begin{aligned} \epsilon^2 \;&=\; |\mathbf{A}\widehat{x}-b|^2 \\ &=\; \left|\left(\mathbf{A}\left[\mathbf{A}^T\mathbf{A}\right]^{-1}\mathbf{A}^T-\mathbf{I}\right)b\right| \\ &\leq\; \left\|\mathbf{A}\left[\mathbf{A}^T\mathbf{A}\right]^{-1}\mathbf{A}^T-\mathbf{I}\right\||b| \end{aligned}$$

where $\|\mathbf{A}\|$ is the matrix norm[1] induced by the Euclidean 2-norm.

## 2. Linear Algebra

**Linear algebra** is a branch of mathematics concerned with the algebraic properties of mathematical systems that generalize our traditional notion of a vector space. The generalization of a vector space is called a *linear space* and mappings between linear spaces are called *linear transformations*. From an engineer's perspective, we view linear spaces as *signal spaces* and linear transformations are used as mathematical models for *systems* that generate these signals. So linear algebra becomes an important formal tool for the modeling, prediction, and design of engineering systems. The following subsections review basic concepts in linear algebra that are used throughout the remaining lectures.

**2.1. Linear Spaces.** A *linear space* generalizes our notion of a *vector space* to objects that are not necessarily vectors. We start with a generalization of *real numbers* into an algebraic system known as a *field*. In particular, a field $F = (X, +, \times)$ is a triple formed from a set $X$ and two binary operations called *addition*, $+$, and *multiplication*, $\times$ that satisfy certain conditions given below. The sum of two elements $x, y \in X$ is denoted as $x + y$ and the product of two elements $x, y \in X$ is denoted as $x \times y$ or $xy$. The conditions that the two binary operations must satisfy are

- Addition $(+)$ is commutative, associative, and *closed* in $X$. There exists an *additive identity* $0 \in X$ and each element $x \in X$ has an *additive inverse*, $-x \in X$.
- Multiplication $(\times)$ is commutative, associative, and closed in $X$. There exists a *multiplicative identity* $1 \in X$ such that $1 \neq 0$ and every nonzero element $x \in X$, has a *multiplicative inverse*, $x^{-1}$.

[1]The matrix norm of a matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is defined as $\|\mathbf{A}\| = \max_{x \neq 0} \frac{|\mathbf{A}x|}{|x|}$ of $\mathbf{A}$.

- Addition and multiplication satisfy the *distributive* laws,

$$\begin{aligned} x(y+z) &= xy + xz \\ (x+y)z &= xz + yz \end{aligned}$$

  for all $x, y, z \in X$.

The set of real numbers is obviously a field. There are other sets of mathematical objects that form a field once a suitable pair of binary operations are chosen. The set of complex numbers, $\mathbb{C}$, forms a field with respect to complex addition and complex multiplication. The set of *rational numbers* (denoted as $\mathbb{Q}$) form a field with respect to their usual binary operations. The set of *rational functions* is formed from the ratio of two polynomials. This set forms a field with respect to polynomial addition and polynomial multiplication.

A *linear space* is formally defined with respect to a set $X$ and a field $F$. The elements of $X$ will be called *vectors* (we are overloading the name "vector" in this case) and the elements of $F$ are called *scalars*. We introduce a binary operation over $X$ called *addition* $(+)$ that maps ordered pairs of vectors $(x, y) \in X \times X$ onto a single vector $x + y \in X$. We also introduce a binary operation called *dilation* $(\cdot)$ that maps a scalar-vector pair $(\alpha, x) \in F \times X$ onto a vector $\alpha x \in X$. We will often denote $\alpha x$ as $\alpha \cdot x$. The ordered tuple, $L = (X, F, +, \cdot)$ is called a *linear space* if

- Addition is commutative, associative, and closed in $X$. There exists an additive identity, $\overline{0} \in X$ and for each $x \in X$ there is an additive inverse, $-x \in X$.
- For all $x \in X$ and $\alpha, \beta \in F$ there exist vectors $\alpha x \in X$ and $\beta x \in X$ such that

$$\alpha \cdot (\beta \cdot x) = \alpha \cdot (\beta x) = \alpha(\beta x) = (\alpha\beta)x$$

  and for all $x \in X$ we have $1 \cdot x = x$ where $1$ is the multiplicative identity of the field, $F$.

- Addition and dilate distribute as

$$\begin{aligned}(\alpha+\beta)x &= (\alpha x)+(\beta y)\\ \alpha(x+y) &= (\alpha x)+(\alpha y)\end{aligned}$$

  for all $x, y \in X$ and $\alpha, \beta \in F$.

Euclidean $n$-space, $\mathbb{R}^n$, is clearly a linear space where $X = \mathbb{R}^n$ and $F = \mathbb{R}$. The set of real and complex valued functions also form a linear space. These spaces are useful since we think of dynamical systems as generating *signals* that are a function of time. This means that the inputs and outputs of a dynamical system are elements of a linear space that we refer to as a *signal space*.

Let us try to justify this idea more formally. Consider a *continuous* real valued function, $x : \mathbb{R} \to \mathbb{R}$. Denote the set of all continuous functions as $C(\mathbb{R}, \mathbb{R})$. For any functions $x, y \in C(\mathbb{R}, \mathbb{R})$, define function addition in a component-wise manner as

$$(x + y)(t) = x(t) + y(t)$$

for all $t \in \mathbb{R}$. Note that on the right hand side of this equation the symbol, $+$, acts like real addition whereas on the left hand side the symbol, $+$, acts on two functions in $C(\mathbb{R}, \mathbb{R})$. We define scalar-dilation in a similar manner in which

$$(\alpha x)(t) = \alpha \cdot x(t)$$

for all $t \in \mathbb{R}$. The fact that addition and dilation on the right hand side of these equations are defined with respect to the vector space, $\mathbb{R}$, means that these new "function" binary operations on the left hand side inherit many of the attributes of vector-space addition and scalar-vector multiplication.

The only property that is not inherited in this manner is *closure*. In other words, we cannot conclude that $C(\mathbb{R}, \mathbb{R})$ is closed with respect to these two binary operations. Closure would mean that the sum of any two continuous functions is continuous and the dilation of any continuous function is also

continuous. Establishing continuity of $x + y$ and $\alpha x$ requires that we use the fundamental definitions for continuity.

Remember that a function $x : \mathbb{R} \to \mathbb{R}$ is continuous if and only if for all $\epsilon > 0$ there exists $\delta > 0$ such that $|x(t) - x(t')| < \epsilon$ whenever $|t - t'| < \delta$. So to establish that $x + y$ is continuous with $x$ and $y$ are continuous, we would need to certify that the above definition for continuity also holds for $x + y$. So let us assume $x$ and $y$ are continuous so for any $\epsilon$ there exist $\delta_1$ and $\delta_2$ such that $|x(t) - x(t')| < \frac{\epsilon}{2}$ when $|t - t'| < \delta_1$ and $|y(t) - y(t')| < \frac{\epsilon}{2}$ when $|t - t'| < \delta_2$. Note that

$$\begin{aligned} |(x + y)(t) - (x + y)(t')| &= |x(t) + y(t) - x(t') - y(t')| \\ &\leq |x(t) - x(t')| + |y(t) - y(t')| \end{aligned}$$

If we select $\delta = \min(\delta_1, \delta_2)$, then we know both terms on the left hand side are less than $\frac{\epsilon}{2}$. This means that when $|t - t'| < \delta$ then

$$|(x + y)(t) - (x + y)(t')| \leq |x(t) - x(t')| + |y(t) - y(t')| \leq \epsilon$$

Since our choice of $\epsilon$ was arbitrary, we have found the required $\delta$ and so $x + y$ is also continuous. A similar argument can also be used for $\alpha x$ and this means that $C(\mathbb{R}, \mathbb{R})$ is closed with respect to the two binary operations. This means for these binary operations, the set $C(\mathbb{R}, \mathbb{R})$ satisfies the axioms of a linear space and so $C(\mathbb{R}, \mathbb{R})$ is a linear space. We sometimes refer to it as a *signal space*.

Let $X$ be a linear space over a field $F$ and let $x_1, \ldots, x_n \in F$ and $\alpha_1, \cdots, \alpha_n \in F$, then the vector

$$x = \sum_{i=1}^{n} \alpha_i x_i$$

is called a *linear combination* of vectors $x_1, \ldots, x_n$. A non-empty subset $M$ of $X$ is called a *subspace* of $X$ if for any pair of scalars $\alpha, \beta \in F$ and any pair of vectors, $x, y \in M$ we have $\alpha x + \beta y \in M$. In other words, a subspace is *closed* under linear combinations of its elements. Consider a a collection, $M = \{x_1, \ldots, x_m\}$ of vectors in linear space, $X$. The set of all

linear combinations formed from elements of $M$ is a subspace called the *span* of $M$. This subspace is denoted as

$$\text{span}(M) \stackrel{\text{def}}{=} \left\{ x \in X \,:\, x = \sum_{i=1}^{m} \alpha_i x_i, \quad \text{where } \alpha_i \in F \text{ and } x_i \in M \right\}$$

Given a collection $M = \{x_1, \ldots, x_m\}$ of vectors in linear space, $X$, if there exists a set of scalars, *not all zero*, such that

$$\sum_{i=1}^{m} \alpha_i x_i = 0$$

then $M$ is said to be *linearly dependent*. This set, $M$, is said to be *linearly independent* if the above equation is only satisfied when all $\alpha_i$ are zero. If $M$ is linearly dependent, then there exist nonzero scalars $\alpha_{i_1}, \alpha_{i_2}, \ldots, \alpha_{i_r}$ with $r < m$ such that

$$\sum_{j=1}^{r} \alpha_{i_j} x_{i_j} = 0 \quad \Rightarrow \quad x_{i_1} = -\frac{1}{\alpha_{i_1}} \sum_{j=2}^{r} \alpha_{i_j} x_{i_j}$$

This means that all vectors in a linearly dependent collection $M$ can be written as a linear combination of other vectors in $M$.

Given a collection $M = \{x_1, \ldots, x_m\}$ of vectors in $X$, we say $M$ forms a *basis* for $X$ if $M$ spans $X$ and $M$ is linearly independent. $X$ is said to be *finite-dimensional* if there exists a basis for $X$ having a finite number of elements. Any basis of a finite dimensional linear space, $X$, has the same number of basis elements that we call the *dimension* of $X$ and denote it as $\dim(X)$.

Let $X$ be a finite-dimensional linear space and let $B = \{e_1, \ldots, e_m\}$ be a basis for $X$. Consider a vector $x \in X$ and assume it has two different linear combinations

$$x = \sum_{i=1}^{m} \alpha_i e_i = \sum_{i=1}^{m} \beta_i e_i$$

If we take the difference of both representations for $x$ we see that

$$\overline{0} = \sum_{i=1}^{m} (\alpha_i - \beta_i) e_i$$

Since $B$ is basis, we know that $\{e_i\}_{i=1}^m$ is a linearly independent collection of vectors. So the only coefficients, $\alpha_i - \beta_i$, satisfying the above equations must be equal to zero. In other words, $\alpha_i = \beta_i$ for $i = 1, 2, \ldots, n$ and this means that every nonzero $x \in X$ has a *unique* representation as a linear combination of its basis set. We can think of the coefficients of $x$ as a *concrete representation* of $x \in X$ with respect to basis $B$. In particular, these coordinates can be arranged as a column vector whose components take values in the field $F$

$$[x]_B = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{bmatrix} \in F^m$$

where the notation $[x]_B$ is used to denote the *concrete representation* of $x$ with respect to basis $B$. Basically what this shows is that any finite dimensional linear space has a concrete representation as a traditional vector space.

A finite-dimensional *linear space*, $L = (X, F, +, \cdot)$, is an *algebraic system* that behaves algebraically like a vector space. Algebraic means that the behavior of the two binary operations (addition and dilation) behave similarly to what we see in vector spaces. However, we also know that vector spaces have a "topological" or "metric" structure that allows one to define how "close" or "similar" two elements of the space are to each other. There is nothing in the "algebraic" definition of a linear space to provide this notion of "closeness". So we will find it convenient to endow our linear spaces with this metric structure and this is done most easily by attaching a function $\|\cdot\| : X \to \mathbb{R}$ called the *norm* on the linear space, thereby turning the linear space into a *normed linear space*.

Intuitively, a *norm* measures the "size" or "length" of an element in the linear space, $X$. So we can formally define what properties the norm function has by abstracting those properties that we usually associate with our

notion of "vector" length. We formally define the norm of $x \in X$ as a real number $\|x\| \in \mathbb{R}$ such that

- $\|x\| \geq 0$ and $\|x\| = 0$ if and only if $x = \overline{0}$
- For any $\alpha \in F$ and $x \in X$, we have $\|\alpha x\| = |\alpha|\|x\|$.
- and for all $x, y \in X$ we have

$$\|x + y\| \leq \|x\| + \|y\|$$

There are several commonly used norms for linear spaces formed from real-valued functions and functions of a complex variable. We review some of these norms and their associated normed linear spaces below.

Consider a linear space $L(\mathbb{R}, \mathbb{R}^n)$ of *integrable continuous-time functions*, $x : \mathbb{R} \to \mathbb{R}^n$. We define the $\mathcal{L}_p$ norm of $x \in L(\mathbb{R}, \mathbb{R}^n)$ where $p$ is a positive integer as

$$\|x\|_{\mathcal{L}_p} \stackrel{\text{def}}{=} \lim_{T \to \infty} \left( \int_{-T}^{T} |x(\tau)|^p d\tau \right)^{1/p}$$

where $|x(\tau)|$ is the Euclidean 2-norm of the vector $x(\tau) \in \mathbb{R}^n$. We define the *normed linear space*, $\mathcal{L}_p$ as the linear space consisting of all functions $x \in L(\mathbb{R}, \mathbb{R}^n)$ such that $\|x\|_{\mathcal{L}_p}$ is finite

$$\mathcal{L}_p \stackrel{\text{def}}{=} \left\{ \text{Integrable functions in } x \in L(\mathbb{R}, \mathbb{R}^n) \text{ such that } \|x\|_{\mathcal{L}_p} = M < \infty \right\}$$

The most commonly used $\mathcal{L}_p$ norms are for $p = 1$, $p = 2$, and $p = \infty$. For $p = \infty$, the norm is

$$\|x\|_{\mathcal{L}_\infty} \stackrel{\text{def}}{=} \lim_{p \to \infty} \|x\|_{\mathcal{L}_p} = \max_i \left\{ \sup_{t \in \mathbb{R}} |x_i(t)| \right\}$$

where $|x_i(t)|$ is the absolute value of the $i$th component of the vector $x(t)$. The $\mathcal{L}_\infty$ space is then the space of all integrable functions with a finite $\mathcal{L}_\infty$ norm.

Another important set of normed linear spaces is generated by the Laplace transform of real-valued functions. In particular, let $H(\mathbb{C}, \mathbb{C})$ denote the linear space of all functions of a complex variable. We define the 2-norm and

$\infty$-norm of a signal $X \in H(\mathbb{C}, \mathbb{C})$ as

$$\begin{aligned} \|X\|_{\mathcal{H}_2} &= \left( \sup_{\alpha>0} \frac{1}{2\pi} \int_{-\infty}^{\infty} |X(\alpha + j\omega)|^2 \, d\omega \right)^{1/2} \\ \|X\|_{\mathcal{H}_\infty} &= \sup_{\alpha>0} \sup_{\omega \in \mathbb{R}} |X(\alpha + j\omega)| \end{aligned}$$

Note that these norms are only defined if the right hand side of the equation is finite. This will only occur if $X$ is *analytic* on the right hand side of the complex plane. A function $X : \mathbb{C} \to \mathbb{C}$ is said to be analytic at a point $z \in \mathbb{C}$ if and only if there is neighborhood about $z$ in which $X(z)$ has a derivative at each point in that neighborhood. This condition essentially means that $X(z)$ has continuous derivatives of all orders and that it can be represented by a convergence power series. Essentially this also means that $z$ cannot be a removable singularity (pole) of $X$. The normed linear spaces associated with these two norms for $H(\mathbb{C}, \mathbb{C})$ consist of those functions with bounded norms, and so they can be written as follows

$$\begin{aligned} \mathcal{H}_2 &= \left\{ X \in H(\mathbb{C}, \mathbb{C}) \; : \; \begin{array}{l} X \text{ is analytic in RHS of complex plane} \\ \text{and } \|X\|_{\mathcal{H}_2} < \infty \end{array} \right\} \\ \mathcal{H}_\infty &= \left\{ X \in H(\mathbb{C}, \mathbb{C}) \; : \; \begin{array}{l} X \text{ is analytic in RHS of complex plane} \\ \text{and } \|X\|_{\mathcal{H}_\infty} < \infty \end{array} \right\} \end{aligned}$$

Note that a norm is not the only way one can establish a topological structure on a linear space. Another commonly used approach is through the introduction of an *inner product*, which yields an inner product space. Inner products generalize the notion of dot products in vector spaces. Recall that the dot product of two vectors, $x, y \in \mathbb{R}^n$ is defined as $x^T y$. When $x$ and $y$ are elements of an abstract linear space, then the corresponding concept is that of an *inner product*. Consider a binary operation $\langle \cdot, \cdot \rangle : X \times X \to \mathbb{C}$ maps a pair of vectors in $x, y \in X$ onto a complex number, $\langle x, y \rangle$. This binary operation is called an *inner product* if

- For all $x, y, z \in X$ we have $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$
- $\langle x, y \rangle = \langle y, x \rangle$

- $\langle x, x\rangle \geq 0$ with $\langle x, x\rangle = 0$ if and only if $x = \overline{0}$.

A linear space equipped with an inner product is called an inner product space. Clearly the dot product we discussed above is an inner product for real-valued vector spaces. If we consider the linear space of integrable functions, $L(\mathbb{R}, \mathbb{R}^n)$, then a commonly used inner product is

$$\langle f, g\rangle = \int_{-\infty}^{\infty} f^T(\tau) g(\tau) d\tau$$

for any $f, g \in L(\mathbb{R}, \mathbb{R}^n)$.

Inner products are used to determine if two elements of a linear space are *orthogonal* to each other. In particular we say $x, y \in X$ are orthogonal if and ony if $\langle x, y\rangle = 0$. Orthogonality and linear independence are related concepts. In particular, if we consider a set of vectors $\{x_1, \ldots, x_m\}$ drawn from linear space, $X$, and we assume they are mutually orthogonal (i.e. $\langle x_i, x_j\rangle = 0$ for all $i \neq j$), then if we consider any linear combination of the form

$$\overline{z} = \sum_{i=1}^{m} \alpha_i x_i \equiv y$$

Taking the inner product of $y$ with any $x_k$ from the collection of orthogonal vectors yields

$$0 = \langle y, x_k\rangle = \sum_{i=1}^{m} \alpha_i \langle x_i, x_k\rangle = \alpha_k \langle x_k, x_k\rangle$$

Since $\langle x_k, x_k\rangle > 0$ and since our choice of $k$ was arbitrary, we can conclude that $\alpha_i = 0$ for all $i = 1, 2, \ldots, m$. This means that the collection $\{x_1, \ldots, x_m\}$ is linearly independent. So we have just proven that a set of mutually orthogonal vectors is also linearly independent.

If one can talk about vectors being orthogonal to each other, we can also speak of subspaces being orthogonal to each other. In particular, one says two linear subspaces, $U$ and $V$, of linear space $X$ are *orthogonal* if every vector in $U$ is orthogonal to every vector in $V$. Given a subspace $U$ of $X$, the

space of all vectors orthogonal to $U$ is called the *orthogonal complement*, $U^{\perp}$, of $U$.

**2.2. Linear Transformations.** A linear transformation is a mapping between elements of two linear spaces. If these linear spaces are real-valued vector spaces, then the linear transformation is a matrix. If the linear spaces consist of real-valued integrable functions, then the linear transformation can be written explicitly as an integral transform, or implicitly as a set of differential equations. To formalize this notion, let $X$ and $Y$ be two linear spaces over the same field, $F$. Let $\mathbf{G} : X \to Y$ denote a function taking elements of $X$ onto elements of $Y$. Let $\mathbf{G}[x]$ denote the element in $Y$ associated with the argument $x \in X$. We say $\mathbf{G}$ is a *linear transformation* if it satisfies the *principle of superposition*. This means that for any $x, y \in X$ and $\alpha, \beta \in F$ we have

$$\mathbf{G}\left[\alpha x + \beta y\right] = \alpha \mathbf{G}[x] + \beta \mathbf{G}[y]$$

It will be convenient to define the *zero*-transformation $\mathbf{0}[x] = \overline{0}$ and the *identity* transformation $\mathbf{I}[x] = x$ for all $x \in X$. These two transformations are easily shown to also be linear transformations.

Let $L(X, Y)$ denote the set of *all* linear transformations from linear space $X$ to linear space $Y$. Define the *addition* of two linear transformations, $\mathbf{G}, \mathbf{H} \in L(X, Y)$ in a component-wise manner. In other words, for any $x \in X$ we have

$$(\mathbf{G} + \mathbf{H})[x] = \mathbf{G}[x] + \mathbf{H}[x]$$

Where $+$ on the right hand side is addition for the linear space, $Y$ and $+$ and the left hand side is addition of the two linear transformations. Define the *dilation* of $\mathbf{G}$ with respect to any $\alpha \in F$ as

$$(\alpha \mathbf{G})[x] = \alpha \left(\mathbf{G}[x]\right)$$

for all $x \in X$. One can also show that $\mathbf{G} + \mathbf{H}$ and $\alpha \mathbf{G}$ are linear transformations so we can assert that $L(X, Y)$ is *closed* under these two binary

operations. Since the range space of these transformations is also a linear space, we can readily conclude that these two binary operations satisfy all of the axioms for a linear space. In other words, the set of all linear transformations, $L(X,Y)$, is also a linear space. In applications, we would think of a linear transformation is a *linear system* which means that $L(X,Y)$ can be thought of as a *linear system space*.

Since $L(X,Y)$ is a linear space, it has a basis. Let $X$ be a finite dimensional linear space with basis set $\{e_1,\ldots,e_n\}$. Let $Y$ be a finite dimensional linear space with basis set $\{f_1,\ldots,f_m\}$. Consider the set of $mn$ linear transformations $\mathbf{E}_{ij}: X \to Y$ (for $i = 1,2,\ldots,m$ and $j = 1,2,\ldots,n$) that takes values

$$\mathbf{E}_{ij}[e_k] = \delta_{jk} f_i$$

where $\delta_{jk}$ is the Kronecker delta

$$\delta_{jk} = \begin{cases} 1 & \text{if } j = k \\ 0 & \text{otherwise} \end{cases}$$

One can show that these $\mathbf{E}_{ij}$ transformations for a basis for the linear space $L(X,Y)$ of linear transformations.

Now consider any linear transformation, $\mathbf{G} \in L(X,Y)$. Since $\{\mathbf{E}_{ij}\}$ is a basis for $L(X,Y)$ we know that for any $k = 1,2,\ldots,n$ there are unique coefficients $\beta_{ik} \in F$ such that

$$\mathbf{G}[e_k] = \sum_{i=1}^{m} \beta_{ik} f_i$$

Let $B_x$ denote the basis $\{e_1, e_2, \ldots, e_n\}$ and let $B_y$ denote the basis $\{f_1, f_2, \ldots, f_m\}$ for $Y$. Note that

$$\begin{aligned} \mathbf{G}[x] &= \mathbf{G}\left[\sum_{j=1}^{n} \alpha_j e_j\right] = \sum_{j=1}^{n} \alpha_j \sum_{i=1}^{m} \beta_{ij} f_i \\ &= \sum_{i=1}^{m} \left(\sum_{j=1}^{n} \beta_{ij}\alpha_j\right) f_i \end{aligned}$$

This is a matrix vector equation that can be written as

$$\begin{aligned}[\mathbf{G}[x]]_{B_y} &= \begin{bmatrix} \sum_{j=1}^{n} \beta_{1j}\alpha_j \\ \vdots \\ \sum_{j=1}^{n} \beta_{mj}\alpha_j \end{bmatrix} = \begin{bmatrix} \beta_{11} & \cdots & \beta_{1n} \\ \vdots & \ddots & \vdots \\ \beta_{m1} & \cdots & \beta_{mn} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} \\ &= [\mathbf{G}]_{B_y}^{B_x} [x]_{B_x}\end{aligned}$$

where the notational convention, $[\mathbf{G}]_{B_y}^{B_x}$, denotes the *concrete representation* of $\mathbf{G}$ with respect to the inputs space's basis, $B_x$, and the output space's basis, $B_y$. The linear transformation, $\mathbf{G}$, is therefore *concretely* represented as a matrix, $[\mathbf{G}]_{B_y}^{B_x}$ once the bases $B_x$ and $B_y$ have been chosen for the linear transformation's input and output spaces. This observation is useful because it means that more complex linear transformations representing dynamical systems can be concretely viewed as matrices, so that many of the geometric intuitions we have when matrices act on vector spaces can also be used to understand what happens when a linear transformation acts on finite-dimensional linear spaces.

Since linear transformations form a linear space, we will also find it useful to introduce a topology or metric on these linear spaces as well. Let us consider a system $\mathbf{G} : \mathcal{L}_2 \to \mathcal{L}_2$ mapping finite energy input signals onto finite energy output signals. We will find it convenient to refer to $\mathbf{G}$ as a *system* mapping $\mathcal{L}_2$ signals onto $\mathcal{L}_2$ signals. The amount of energy gained or lost between the input and output is sometimes called a *gain* for the system. We can therefore define the system's $\mathcal{L}_2$-induced gain as

$$\begin{aligned}\|\mathbf{G}\|_{\mathcal{L}_2-\text{ind}} &\stackrel{\text{def}}{=} \sup_{w\neq 0} \frac{\|\mathbf{G}[w]\|_{\mathcal{L}_2}}{\|w\|_{\mathcal{L}_2}} \\ &= \sup_{\|w\|_{\mathcal{L}_2}=1} \|\mathbf{G}[w]\|_{\mathcal{L}_2}\end{aligned}$$

In other words, the system's (linear transformation's) $\mathcal{L}_2$ induced gain equals the largest output energy seen over all applied inputs with unit energy. This means that the actual output energy of the system satisfies the inequality

$$\|\mathbf{G}[w]\|_{\mathcal{L}_2} \leq \|\mathbf{G}\|_{\mathcal{L}_2-\text{ind}} \|w\|_{\mathcal{L}_2}$$

The $\sup$ in the above definition means there is a specific signal, $w$, for which the above inequality holds with equality. We may, therefore, obtain an equivalent characterization of the induced gain as

$$\|\mathbf{G}\|_{\mathcal{L}_2-\text{ind}} = \inf\left\{\gamma \in \mathbb{R} \,:\, \|\mathbf{G}[w]\|_{\mathcal{L}_2} \leq \gamma \|w\|_{\mathcal{L}_2}\right\}$$

We say $\|\mathbf{G}\|_{\mathcal{L}_2-\text{ind}}$ is an *induced* gain because it is "induced" by our selection of the norms for the input and output signal spaces. In this case, we chose the $\mathcal{L}_2$ on both spaces, hence the name $\mathcal{L}_2$-induced gain. It may be more convenient from the application's standpoint to select a different norm on the input and output spaces. For instance many mechanical engineers might prefer to use an $\mathcal{L}_\infty$ norm, in which case, our induced gain would be different.

The formal definition for the induced gain is awkward to work with for it provides no explicit formula we can use to compute what the gain might be. For certain selections of the signal space norms, however, we can find explicit formulas. For example, let $\mathbf{G} : \mathcal{L}_\infty \to \mathcal{L}_\infty$ be a linear transformation defined explicitly through the convolution equation

$$\mathbf{G}[w](t) = y(t) = \int_{-\infty}^{t} g(t-\tau)w(\tau)d\tau$$

where $g : \mathbb{R} \to \mathbb{R}$ is a function called the system's *impulse response function* and $w \in \mathcal{L}_\infty$ is the applied input. To determine the $\mathcal{L}_\infty$ induced gain, we need to bound the $\mathcal{L}_\infty$ norm of the output, so

$$\begin{aligned} |y(t)| &= \left| \int_{-\infty}^{\infty} g(\tau)w(t-\tau)d\tau \right| \\ &\leq \int_{-\infty}^{\infty} |g(\tau)|\,|w(t-\tau)|d\tau \\ &\leq \left[\int_{-\infty}^{\infty} |g(\tau)|d\tau\right] \|w\|_{\mathcal{L}_\infty} = \|g\|_{\mathcal{L}_1} \|w\|_{\mathcal{L}_\infty} \end{aligned}$$

since $\|y\|_{\mathcal{L}_\infty}$ is the largest $|y(t)|$, we have

$$\|y\|_{\mathcal{L}_\infty} = \sup_t \|y(t)\| \leq \|g\|_{\mathcal{L}_1} \|w\|_{\mathcal{L}_\infty}$$

So we can conclude that the $\mathcal{L}_1$ norm of the impulse response function, $g$, is an upper bound on the system's $\mathcal{L}_\infty$ induced gain.

One may show that $\|g\|_{\mathcal{L}_1}$ equals the $\mathcal{L}_\infty$ induced gain by finding any $\mathcal{L}_\infty$ input for which the inequality holds with equality. Finding such "signals" requires some degree in ingenuity, but for this example, if we consider

$$w(t-\tau) = \operatorname{sgn}(g(\tau))$$

it is easy to show that $y(t) = \|g\|_{\mathcal{L}_1}$, thereby establishing that $\|\mathbf{G}\|_{\mathcal{L}_\infty - \text{ind}}$ is equal to $\|g\|_{\mathcal{L}_1}$. Since this signal norm can be explicitly computed once we know the impulse response function, $g$, it provides a concrete way of computing what this system's $\mathcal{L}_\infty$-induced gain.

Consider a linear transformation $\mathbf{G} : X \to Y$ where $X$ and $Y$ are inner product spaces. We will find it convenient to define another linear transformation $\mathbf{G}^* : Y \to X$ that takes vectors in $Y$ onto vectors in $X$. This linear transformation $\mathbf{G}^*$ is called the *adjoint* of $\mathbf{G}$ if for any $x \in X$ and $y \in Y$ we have

$$\langle \mathbf{G}[x], y \rangle = \langle x, \mathbf{G}^*[y] \rangle$$

Adjoints are useful technical tools for sometimes it is easier to establish results regarding a linear transformation using the adjoint, rather than the original transformation itself.

Consider a linear transformation $\mathbf{G} : X \to Y$ between two linear spaces $X$ and $Y$. This linear transformation has two important subspaces; its null space and its range space. The null space of $\mathbf{G}$ (also known as the linear transformation's *kernel*) is

$$\ker(\mathbf{G}) \stackrel{\text{def}}{=} \{x \in X \, : \, \mathbf{G}[x] = 0\}$$

The *range space* of $\mathbf{G}$ is

$$\text{Range}(\mathbf{G}) \stackrel{\text{def}}{=} \{y \in Y \, : \, \text{there exists } z \in X \text{ such that } y = \mathbf{G}[z]\}$$

Recall that when $X$ and $Y$ are finite dimensional then any element $x \in X$ and $y \in Y$ has a concrete representation with respect to the field $F$ once we've chosen a basis for $X$ and $Y$. Let $B_x$ denote the basis for $X$ and $B_y$ denote the basis of $Y$. The concrete representations for $x$ and $y$ (following our earlier notational convention) are $[x]_{B_x} \in F^n$ and $[y]_{B_y} \in F^m$, respectively. We also know that $\mathbf{G} : X \to Y$ has a concrete representation as the matrix $[\mathbf{G}]_{B_y}^{B_x} \in F^{m \times n}$ with

$$[y]_{B_y} = [\mathbf{G}]_{B_y}^{B_x} [x]_{B_x}$$

The preceding concrete representation is a linear algebraic equation that we studied in the preceding section. Based on those results we know that $y = \mathbf{G}[x]$ has a solution if and only if its concrete representation has a solution. We know that such solutions exist if $y \in \text{Range}\left([\mathbf{G}]_{B_y}^{B_x}\right)$ and this solution is unique if and only if $\ker\left([\mathbf{G}]_{B_y}^{B_x}\right)$ is trivial.

A useful relationship can be established between the null space and range space of a linear transformation $\mathbf{G}$. In particular, this relationship is sometimes called the *fundamental theorem of linear algebra* [Strang (1976)] and it asserts that

$$\ker(\mathbf{G}) = \text{Range}(\mathbf{G}^*)^\perp$$

where $\mathbf{G}^*$ is the adjoint operator of $\mathbf{G}$. To prove this equivalence, suppose that $x \in \ker(\mathbf{G})$ and $y \in \text{Range}(\mathbf{G}^*)$. This would mean $\mathbf{G}[x] = 0$ and for some $z \in Y$ we have $\mathbf{G}^*[z] = y$. Taking the inner product of $y$ and $x$ yields,

$$\langle y, x \rangle = \langle \mathbf{G}^*[z], x \rangle = \langle z, \mathbf{G}[x] \rangle = \langle z, 0 \rangle = 0$$

This shows that any vector $y \in \text{Range}(\mathbf{G}^*)$ is orthogonal to any vector $x \in \ker(\mathbf{G})$. This fundamental theorem of linear algebra provides a useful tool for proving results, where proving a given assertion is easier to do on the adjoint.

**2.3. Eigenvalues and Eigenvectors.** A useful way of characterizing a linear transformation $\mathbf{G} : X \to X$ over a linear space $X$ is to determine

those vectors in $X$ that are *invariant*. In other words, we look for those vectors, $v \in X$, that are left "unchanged" by the application of the linear transformation, i.e. $\mathbf{G}[v] = v$. This is actually too restrictive to be useful, so we look for vectors that are "invariant" up to a scalar dilation. So this means we look for ordered pairs, $(v, \lambda) \in X \times F$ where $v$ is a nonzero vector in $X$ and $\lambda$ is a scalar in the field $F$ such that

$$\mathbf{G}[v] = \lambda v$$

The scalar $\lambda$ is called an eigenvalue of $\mathbf{G}$ and $v$ is called its associated *eigenvector*.

When $X$ is finite dimensional, then we know that once we choose a basis, $B$, for $X$, we can concretely represent $\mathbf{G}$ as a matrix. If $X$ is an $n$-dimensional linear space, then $[\mathbf{G}]_B^B \in F^{n\times n}$ is a square matrix whose components are in the field $F$. In many of our applications $F$ will either be the real, $\mathbb{R}$ or complex, $\mathbb{C}$, field. In this case, we can specialize our notion of an eigenvector/value to correspond to a matrix. So given a matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$, then the pair $(v, \lambda) \in \mathbb{C}^n \times \mathbb{C}$ is a *right* eigenvalue/vector pair for $\mathbf{A}$ if

$$\mathbf{A}v = \lambda v$$

We call $(v, \lambda)$ a left eigenvalue vector pair if

$$v^T\mathbf{A} = \lambda v^T$$

Eigenvalue/vector problems arise in a number of ways, but one way that is often taught to undergraduate engineers involves using them to characterize solutions to constant coefficient ordinary differential equations (ODE). We will use an example to review how this is done. Consider the following ODE with given initial condition.

$$\begin{aligned} \dot{x}_1(t) &= 4x_1(t) - 5x_2(t), \quad x_1(0) = 8 \\ \dot{x}_2(t) &= 2x_1(t) - 3x_2(t), \quad x_2(0) = 5 \end{aligned}$$

The initial value problem (IVP) is to find explicit representations for the two functions $x_1 : \mathbb{R}_{\geq 0} \to \mathbb{R}$ and $x_2 : \mathbb{R}_{\geq 0} \to \mathbb{R}$ that satisfy the differential equation for all $t \geq 0$ and that satisfy the given initial condition at time $t = 0$.

Note that the IVP can be rewritten in matrix-vector form as

$$\dot{x}(t) \quad = \quad \mathbf{A}x(t), \quad x(0) = x_0$$

where

$$\mathbf{A} = \begin{bmatrix} 4 & -5 \\ 2 & -3 \end{bmatrix}, \quad x(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}, \quad x_0 = \begin{bmatrix} 8 \\ 5 \end{bmatrix}$$

Let us assume that $x_1$ and $x_2$ are both exponential functions of time of the form

$$\begin{aligned} x_1(t) &= e^{\lambda t} x_{10} \\ x_2(t) &= e^{\lambda t} x_{20} \end{aligned}$$

where $x_0 = \begin{bmatrix} x_{10} \\ x_{20} \end{bmatrix}$ is the initial condition and $\lambda \in \mathbb{C}$ is a complex valued constant. We can substitute this assumed form for the solution into the differential equation to obtain

$$\begin{aligned} \lambda e^{\lambda t} x_{10} &= 4e^{\lambda t} x_{10} - 5e^{\lambda t} x_{20} \\ \lambda e^{\lambda t} x_{20} &= 2e^{\lambda t} x_{10} - 3e^{\lambda t} x_{20} \end{aligned}$$

which we can also write in matrix-vector form as

$$\lambda x_0 = \lambda \begin{bmatrix} x_{10} \\ x_{20} \end{bmatrix} = \begin{bmatrix} 4 & -5 \\ 2 & -3 \end{bmatrix} \begin{bmatrix} x_{10} \\ x_{20} \end{bmatrix} = \mathbf{A}x_0$$

This is a system of linear algebraic equations (LAE) rather than a differential equation and so we can look for $(x_0, \lambda)$ that satisfy this relation

$$\mathbf{A}x_0 = \lambda x_0$$

using the computational procedures discussed in the first section. If a solution does exist, then we know that our guess for the solution was correct

with the exponent $\lambda$ being obtained from the solution of the above set of LAEs. Note, however, that this equation also has the form as the matrix eigenequation we introduced above. So if a solution does exist it means that $\lambda$ will be an eigenvalue of the matrix $\mathbf{A} = \begin{bmatrix} 4 & -5 \\ 2 & -3 \end{bmatrix}$ with associated right eigenvector $x_0$.

How do we find eigenvector/value pairs for a square matrix? Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be a square real-valued matrix and let $v$ be a right eigenvector corresponding to eigenvalue $\lambda \in \mathbb{C}$. This means $\mathbf{A}v = \lambda v$ and $v$ is a nonzero vector. This is equivalent to saying that

$$(\mathbf{A} - \lambda \mathbf{I})v = 0$$

and so it means that the eigenvector $v$ is a nonzero vector in the null space $\ker(\mathbf{A} - \lambda \mathbf{I})$. We know that this homogeneous LAE will have nonzero solutions when the null space of the matrix $\mathbf{A} - \lambda \mathbf{I}$ is nontrivial.

The condition we use to check if a matrix $\mathbf{A}$ has a nontrivial kernel is based on the determinant of that matrix. Given a matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, the determinant of $\mathbf{A}$ (denoted as $\det(\mathbf{A})$) is the sum

$$\det(\mathbf{A}) = \sum_{\sigma \in P(N)} \operatorname{sgn}(\sigma) \cdot a_{ij_1} \cdot a_{2j_2} \cdots a_{nj_n}$$

where $P(N)$ is the set of all possible permutations of $N = \{1, 2, \ldots, n\}$ and for any $\sigma \in P(N)$ we have

$$\operatorname{sgn}(\sigma) = \begin{cases} +1 & \text{if } \sigma \text{ is an even number of permutations from } N \\ -1 & \text{if } \sigma \text{ is an odd number of permutations from } N \end{cases}$$

The main property of the determinant that we use is that the columns (or rows) of $\mathbf{A}$ are linearly dependent if and only if $\det(\mathbf{A}) = 0$.

The computation of $\det(\mathbf{A})$ is easily done using computer tools in MATLAB, so we can use the determinant test to see whether $(\mathbf{A} - \lambda I)$ has a nontrivial null space. For convenience define

$$\mathbf{R}_\lambda = \mathbf{A} - \lambda \mathbf{I}$$

with $i, j$th component, $r_{ij}$. The null space of $\mathbf{R}_\lambda$ is nontrivial if and only if there is a nonzero vector $v \in \mathbb{C}^n$ such that $\mathbf{R}_\lambda v = 0$. We may rewrite this a s

$$0 = \mathbf{R}_\lambda v = v_1 \begin{bmatrix} r_{11} \\ r_{21} \\ \vdots \\ r_{n1} \end{bmatrix} + v_2 \begin{bmatrix} r_{12} \\ r_{22} \\ \vdots \\ r_{n2} \end{bmatrix} + \cdots + v_n \begin{bmatrix} r_{1n} \\ r_{2n} \\ \vdots \\ r_{nn} \end{bmatrix}$$

If $\ker(\mathbf{R}_\lambda)$ is nontrivial then there exists $v \in \ker(\mathbf{R}_\lambda)$ that is nonzero and from the preceding equation this implies the columns of $\mathbf{R}_\lambda$ are linearly dependent. So since $\det(\mathbf{R}_\lambda) = 0$ if and only if the columns of $\mathbf{R}_\lambda$ are linearly dependent, the preceding discussion shows that $\mathbf{A} - \lambda\mathbf{I}$ has a nontrivial null space if and only if $\lambda \in \mathbb{C}$ is chosen so that $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$.

The condition $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$ is actually a polynomial equation with respect to the indeterminate variable $\lambda$. So it will be convenient to define the *characteristic polynomial* of a square matrix $\mathbf{A}$ as

$$p(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I})$$

The condition for $\lambda$ to be an eigenvalue is therefore that $\lambda \in \mathbb{C}$ is a *root* of the matrix characteristic equation

$$p(\lambda) = 0$$

These roots are in the complex field $\mathbb{C}$. If $\lambda_i \in \mathbb{C}$ for $i = 1, 2 \ldots, p$ where $p \leq n$ are *distinct* roots (i.e. all different) of the characteristic equation, then we can factor the characteristic polynomial as

$$p(\lambda) = (\lambda_1 - \lambda)^{m_1}(\lambda_2 - \lambda)^{m_2} \cdots (\lambda_p - \lambda)^{m_p}$$

where $m_i$ is called the *algebraic multiplicity* of the $i$th distinct root, $\lambda_i$.

Note that each distinct eigenvalue of $\mathbf{A}$ has an eigenvector associated with it. If eigenvalue $\lambda$ of $\mathbf{A}$ has an algebraic multiplicity $m > 1$, then the number of linearly independent eigenvectors associated with this eigenvalue will be $\mu \leq m$. The number of linear independent eigenvectors associated

with a distinct eigenvalue $\lambda$ of $\mathbf{A}$ is called the eigenvalue's *geometric multiplicity*.

So let us return to our ODE example and determine the eigenvalues of the matrix $\mathbf{A} = \begin{bmatrix} 4 & -5 \\ 2 & -3 \end{bmatrix}$. We first form the characteristic polynomial

$$\begin{aligned} \det(\mathbf{A} - \lambda \mathbf{I}) &= \det \begin{bmatrix} 4-\lambda & -5 \\ 2 & -3-\lambda \end{bmatrix} \\ &= (4-\lambda)(-3-\lambda) + 10 \\ &= (\lambda + 1)(\lambda - 2) \end{aligned}$$

So this matrix has two distinct eigenvalues $\lambda_1 = -1$ and $\lambda_2 = 2$ that correspond to the roots of $(\lambda + 1)(\lambda - 2) = 0$. Note that each eigenvalue has an algebraic and geometric multiplicity of one, so there is a single nonzero eigenvector associated to each of these distinct eigenvalues.

The eigenvectors are obtained by solving the following system of linear algebraic equations

$$0 = \begin{bmatrix} 4-\lambda & -5 \\ 2 & -3-\lambda \end{bmatrix} x$$

for $x$ using $\lambda = -1$ or $\lambda = 2$. If we let $\lambda = -1$ (the first eigenvalue) we get

$$0 = \begin{bmatrix} 5 & -5 \\ 2 & -2 \end{bmatrix} x \quad \Rightarrow \quad x = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

for $\lambda = 2$ we get

$$0 = \begin{bmatrix} 2 & -5 \\ 2 & -5 \end{bmatrix} x \quad \Rightarrow \quad x = \begin{bmatrix} 5 \\ 2 \end{bmatrix}$$

All solutions of the ODE are obtained by taking a linear combination of these eigensolutions that we just computed. So we have

$$x(t) = c_1 e^{-t} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + c_2 e^{2t} \begin{bmatrix} 5 \\ 2 \end{bmatrix}$$

Our initial condition is $x_0 = \begin{bmatrix} 8 \\ 5 \end{bmatrix}$ and we need to pick $c_1$ and $c_2$ so that $x(0) = x_0$ when $t = 0$. This also gives rise to an LAE of the form

$$\begin{bmatrix} 1 & 5 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} 8 \\ 5 \end{bmatrix}$$

The solution to this LAE is $c_1 = 3$ and $c_2 = 1$, so the full solution to our original IVP is

$$x(t) = 3e^{-t} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + e^{2t} \begin{bmatrix} 5 \\ 2 \end{bmatrix}$$

for all $t \geq 0$.

Consider a linear space, $X$ and consider the two linear transformations $\mathbf{T}_1, \mathbf{T}_2 \in L(X, X)$. Assume there exists an invertible linear transformation $\mathbf{Q} \in L(X, X)$ such that

$$\mathbf{Q}\mathbf{T}_1 = \mathbf{T}_2\mathbf{Q}$$

Let $v$ be an eigenvector of $\mathbf{T}_2$ with eigenvalues $\lambda$, then

$$\begin{aligned} \mathbf{T}_2 v = \lambda v \qquad &\Rightarrow \qquad \mathbf{Q}\mathbf{T}_1(\mathbf{Q}^{-1}v) = \lambda v \\ &\Rightarrow \qquad \mathbf{T}_1(\mathbf{Q}^{-1}v) = \lambda \mathbf{Q}^{-1}v \end{aligned}$$

This last relation says that the vector $\mathbf{Q}^{-1}v$ is an eigenvector of $\mathbf{T}_1$ with the *same* eigenvalue $\lambda$. The linear transformations $\mathbf{T}_1$ and $\mathbf{T}_2$ therefore have the same eigenvalues and their eigenvectors are related through an invertible coordinate transformation, $\mathbf{Q}$. We say that the matrices $\mathbf{T}_1$ and $\mathbf{T}_2$ are *similar* and we refer to $\mathbf{Q}$ as a *similarity transformation*.

Similarity transformations provide a useful way of transforming a matrix into a similar form that is easier to work with. Such "convenient" forms are called *canonical forms*. One important canonical form of a matrix is its *diagonal form* (when it exists) that is also known as its *modal form*. In particular, consider a square matrix $\mathbf{A}$ with $n$ *distinct* eigenvalues $\{\lambda_1, \ldots, \lambda_n\}$

with associated eigenvectors $\{v_1, \ldots, v_n\}$. Define the matrices

$$\mathbf{V} = \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix}, \quad \mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_n) = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{bmatrix}$$

We can therefore see that

$$\begin{aligned} \mathbf{AV} &= \mathbf{A}\begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix} = \begin{bmatrix} \lambda_1 v_1 & \cdots & \lambda_n v_n \end{bmatrix} \\ &= \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix} \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{bmatrix} \\ &= \mathbf{V\Lambda} \end{aligned}$$

So we can see that $\mathbf{V}$ is a similarity transformation matrix between $\mathbf{A}$ and $\mathbf{\Lambda}$, where $\mathbf{\Lambda}$ is a diagonal matrix whose diagonal elements are the eigenvalues.

So if we take our original ODE

$$\dot{x} = \mathbf{A}x, \quad x = x_0$$

and introduce the similarity transformation $\mathbf{V}y = x$ then we get

$$\mathbf{V}\dot{y} = \mathbf{AV}y, \quad \mathbf{V}y_0 = x_0$$

which implies that

$$\dot{y} = \mathbf{V}^{-1}\mathbf{AV}y = \mathbf{\Lambda}y, \quad y_0 = \mathbf{V}^{-1}x_0$$

This is "similar" to the original system. In fact it is better to say that the two systems are *topologically equivalent* for the trajectories of both ODEs can be mapped into each other through the invertible matrix $\mathbf{V}$. Note that this diagonalized system is "easy" to solve since both components are decoupled. In particular, we can see that

$$y_1(t) = c_1 e^{-t}, \quad y_2(t) = c_2 e^{2t}$$

where $c_1$ and $c_2$ are chosen to satisfy the initial condition $y(0) = \mathbf{V}^{-1}x_0$. We can then recover our earlier solution in the original coordinate frame through the equation $x(t) = \mathbf{V}^{-1}y(t)$.

Note that in the preceding example we assumed that $\mathbf{A}$ had $n$ distinct eigenvalues. This is a sufficient condition for $\mathbf{A}$ to be diagonalizable through a similarity transformation. In general, however, this may not always be the case. If some eigenvalues have a geometric multiplicity that is less than the eigenvalue's algebraic multiplicity, then it is not possible to diagonalize the matrix. The best we can do is convert it to its *Jordan Canonical Form*.

Let the characteristic polynomial of $\mathbf{A}$ be

$$p(\lambda) = (\lambda_1 - \lambda)^{m_1} \cdots (\lambda_p - \lambda)^{m_p}$$

such that $\sum_{i=1}^{p} m_i = n$. The Jordan Canonical form of $\mathbf{A}$ is

$$\mathbf{J} = \begin{bmatrix} \mathbf{J}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{J}_p \end{bmatrix}$$

where $\mathbf{J}_i$ is an $m_i \times m_i$ matrix of the form

$$\mathbf{J}_i = \lambda_i \mathbf{I}_{m_i} + \mathbf{N}_i$$

in which

$$\mathbf{N}_i = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix}$$

The Jordan canonical form always exists.

**2.4. PCA and Singular Value Decompositions.** Principal component analysis (PCA) is a useful way for taking a set of data points $\mathcal{D} = \{x_k\}_{k=1}^M$ where $x_k \in \mathbb{R}^n$ and identifying a set of orthogonal vectors $\mathcal{P} = \{p_k\}_{k=1}^m$ whose span form a subspace that minimizes the mean squared error between vectors in the subspace $\mathcal{P}$ and the datapoints in $\mathcal{D}$. This problem has important applications in data compression and machine learning applications.

Let the data samples $\mathcal{D} = \{x_k\}_{k=1}^M$ be real valued vectors in $\mathbb{R}^n$ and let us construct a *data matrix*

$$\mathbf{X} = \begin{bmatrix} x_1 & x_2 & \cdots & x_M \end{bmatrix}$$

whose columns are the data sample vectors, $x_k$. This matrix, therefore, lies in $\mathbb{R}^{n\times M}$. Now let $\mathbf{P}$ be a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$ where $m < n$. We let $\mathbf{P}$ be a concrete representation of this linear transformation of the form

$$\mathbf{P} = \begin{bmatrix} p_1^T \\ p_2^T \\ \vdots \\ p_m^T \end{bmatrix}$$

where $p_k \in \mathbb{R}^n$. If we then consider a new data matrix obtained by transforming $\mathbf{X}$ through $\mathbf{Y}$,

$$\mathbf{Y} = \mathbf{P}\mathbf{X}$$

we see that $\mathbf{Y} \in \mathbb{R}^{m\times M}$ is a matrix whose columns are the projections of $x_k$ onto a *lower dimensional* (remember $m < n$) *latent space*. Essentially, we can think of $\mathbf{Y}$ as a lower dimensional representation for the information in the original data sample, $\mathbf{X}$. Since it is lower dimensional, we have essentially "compressed" the original data vectors in $\mathbf{X}$ into a lower dimensional vector $y \in \mathbf{Y}$.

The rows of $\mathbf{P}$ are said to be *principal components* of $\mathbf{X}$ if they maximize the trace of the covariance matrix of $\mathbf{Y}$

$$\mathbf{C}_\mathbf{Y} = \frac{1}{M}\mathbf{Y}\mathbf{Y}^T$$

subject to $\mathbf{C_Y}$ being diagonal and $\mathbf{PP}^T = \mathbf{I}$. These last two conditions requires that principal components are orthogonal to each other and that they have unit length.

We will now show that the principal components in $\mathbf{P}$ are the $m$ eigenvectors of $\mathbf{XX}^T$ that have the largest eigenvalues. Note that the covariance matrix of $\mathbf{Y}$ may be written as

$$\mathbf{C_Y} = \frac{1}{M}\mathbf{YY}^T = \mathbf{P}\left(\frac{1}{M}\mathbf{XX}^T\right)\mathbf{P}^T = \mathbf{PC_X P}^T$$

where $\mathbf{C_X} = \frac{1}{M}\mathbf{XX}^T$ is the covariance matrix of the original data matrix $\mathbf{X}$. We may decompose $\mathbf{XX}^T$ as $\mathbf{V\Lambda V}^T$ where $\mathbf{\Lambda}$ is a diagonal matrix consisting of the eigenvalues of $\mathbf{C_X}$ and $\mathbf{V}$ is a matrix of eigenvectors of $\mathbf{C_X}$ arranged as columns.

Let us choose $\mathbf{P}$ to be a matrix whose rows are eigenvectors of $\mathbf{C_X}$. To simplify this discussion we'll assume that the eigenvalues of $\mathbf{C_X}$ are distinct with an algebraic multiplicity of 1. We can then write the covariance of $\mathbf{Y}$

$$\mathbf{C_Y} = \mathbf{V}^T\mathbf{V\Lambda V}^T\mathbf{V} = \mathbf{\Lambda}$$

where the last line holds because the eigenvectors of a real symmetric matrix are mutually orthogonal. We have just shown that if we choose the rows of $\mathbf{P}$ from the eigenvectors of $\mathbf{C_X}$, then we diagonalize the covariance matrix of $\mathbf{Y}$. That diagonal contains $m$ of the eigenvalues of $\mathbf{C_X}$ which we know must all be positive since $\mathbf{XX}^T$ is symmetric. Clearly we can maximize the trace of $\mathbf{C_Y}$ if we simply form $\mathbf{P}$ from the eigenvector associated with the $m$ largest eigenvalues of $\mathbf{C_X}$.

It is common to use a particular matrix decomposition known as the singular value decomposition (SVD) to compute the principal components. The algorithms used to compute SVDs represent one of the most numerically stable ways of determining the rank of a matrix, especially for very large data matrices. SVDs are also useful in representing the frequency response of MIMO LTI systems and they can be used to characterize how

close a matrix is to be singular. For any $m \times p$ matrix, $\mathbf{Q}$, one can prove that there exist $m \times m$ and $p \times p$ unitary matrices $\mathbf{U}$ and $\mathbf{V}$ and a real $r \times r$ diagonal matrix $\mathbf{\Sigma}$ such that

$$\mathbf{Q} = \mathbf{U} \begin{bmatrix} \mathbf{\Sigma} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{V}^T$$

The matrix $\mathbf{\Sigma}$ has the form

$$\mathbf{\Sigma} = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r)$$

where $\sigma_i \geq \sigma_{i+1}$ for $i = 1, \ldots, r-1$ and $r \leq \min(m, p)$ is the rank of matrix $\mathbf{Q}$. The triple, $(\mathbf{U}, \mathbf{\Sigma}, \mathbf{V})$ is called the *singular value decomposition* of $\mathbf{Q}$. This decomposition is unique and $\sigma_1$ to $\sigma_r$ are called the non-zero singular values of $\mathbf{Q}$. It can be readily shown that these non-zero singular values are also the positive roots of the non-zero eigenvalues of $\mathbf{Q}^T\mathbf{Q}$. The SVD of $\mathbf{Q}$ may also be written as

$$\mathbf{Q} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \sum_{i=1}^{r} \sigma_i u_i v_i^T$$

where $u_i$ and $v_i$ are the $i$th rows of $\mathbf{U}$ and $\mathbf{V}$, respectively.

To see how this relates back to PCA, let us consider a data matrix $\mathbf{X}$ whose columns are the data sample vectors that have been centered with respect to the dataset's mean. Recall that $\mathbf{C} = \frac{1}{M}\mathbf{X}\mathbf{X}^T$ is the covariance matrix of the data matrix. We know the principal components are the eigenvectors of $\mathbf{C}_\mathbf{X}$. Now consider the SVD of the data matrix $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$. Let us express the covariance matrix of $\mathbf{X}$ in terms of its SVD

$$\mathbf{X}\mathbf{X}^T = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\mathbf{V}\mathbf{\Sigma}\mathbf{U}^T = \mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^T$$

We can therefore conclude that

$$\mathbf{C}_\mathbf{X} = \frac{1}{M}\mathbf{U}\mathbf{\Sigma}\mathbf{U}^T = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$$

where $\mathbf{\Lambda}$ is a diagonal matrix whose diagonal elements are $\lambda_i = \frac{\sigma_i^2}{M}$. Since $\mathbf{U}$ is a unitary matrix (i.e. $\mathbf{U}^T\mathbf{U} = \mathbf{I}$) we can readily see that

$$\mathbf{C}_\mathbf{X}\mathbf{U} = \mathbf{U}\mathbf{\Lambda}$$

This means that the columns of $\mathbf{U}$ are the prinicipal components. Since we defined the PCA transformation $\mathbf{P}$ so its rows were the principal component vectors, we have $\mathbf{P} = \mathbf{U}^T$. If we then look at transforming all data points into the PCA coordinates we have

$$\mathbf{Y} = \mathbf{P}\mathbf{X} = \mathbf{U}^T\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \mathbf{\Sigma}\mathbf{V}^T$$

**2.5. Cayley-Hamilton Theorem.** The characteristic polynomial of a matrix $\mathbf{A}$ can be used in other ways. Let us write the characteristic polynomial as, $p(s) = \det(s\mathbf{I} - \mathbf{A})$, with respect to the indeterminate variable $s$. This polynomial may be written as

$$p(s) = a_0 s^n + a_1 s^{n-1} + \cdots + a_{n-1}s + a_n$$

where $a_i$ for $i = 0, 1, \ldots, n$ are real coefficients. We define the matrix function

$$p(\mathbf{A}) := a_0\mathbf{A}^n + a_1\mathbf{A}^{n-1} + \cdots + a_{n-1}\mathbf{A} + a_n\mathbf{I}$$

An important fact about this matrix function is that $p(\mathbf{A}) = 0$ when $p(s)$ is the characteristic polynomial of $\mathbf{A}$. This result is known as the *Cayley-Hamilton Theorem* [Strang (1976)]. It will be useful in our later work. The proof of this relationship will also be useful later.

Let $\mathbf{N}(s)$ be the *classical adjoint*[2] of $\mathbf{A} - s\mathbf{I}$. This matrix is a function of $s$ and may be written as

$$\mathbf{N}(s) = \mathbf{N}_1 s^{n-1} + \mathbf{N}_2 s^{n-2} + \cdots + \mathbf{N}_{n-1}s + \mathbf{N}_n$$

[2]The classical adjoint, $\mathrm{adj}(\mathbf{A})$ of a matrix, $\mathbf{A}$, is obtained by taking the transpose of its cofactor matrix, $\mathrm{cof}(\mathbf{A})$. The cofactor matrix of a square matrix $\mathbf{A}$ is a matrix whose elements are the cofactors of $\mathbf{A}$. The $ij$th cofactor of $\mathbf{A}$ is the determinant of a matrix obtained by deleting the $i$th row and $j$th column of the matrix and multiplying by $-1$ if $i+j$ is odd. The inverse of a matrix $\mathbf{A}$ can be computed from its adjoint and its determinant, $\mathbf{A}^{-1} = [\mathrm{adj}(\mathbf{A})]\,/\det(\mathbf{A})$.

where the $\mathbf{N}_i$ ($i = 1, 2, \ldots, n$) are real valued matrices. This matrix satisfies the equation

$$(\mathbf{A} - s\mathbf{I})\mathbf{N}(s) = \det(\mathbf{A} - s\mathbf{I})\mathbf{I}$$

Expanding this out gives

$$(\mathbf{A} - s\mathbf{I})(\mathbf{N}_1 s^{n-1} + \cdots + \mathbf{N}_{n-1}s + \mathbf{N}_n) = (a_0 s^n + \cdots + a_{n-1}s + a_n)\mathbf{I}$$

Multiplying out the left hand side of the above equation and equating like power yields,

$$\begin{aligned} -\mathbf{N}_1 &= a_0\mathbf{I} \\ \mathbf{A}\mathbf{N}_1 - \mathbf{N}_2 &= a_1\mathbf{I} \\ &\vdots \\ \mathbf{A}\mathbf{N}_{n-1} - \mathbf{N}_n &= a_{n-1}\mathbf{I} \\ \mathbf{A}\mathbf{N}_n &= a_n\mathbf{I} \end{aligned}$$

Multiply the first equation by $\mathbf{A}^n$, the second by $\mathbf{A}^{n-1}$, and so on to obtain

$$\begin{aligned} -\mathbf{A}^n\mathbf{N}_1 &= a_0\mathbf{A}^n \\ \mathbf{A}^n\mathbf{N}_1 - \mathbf{A}^{n-1}\mathbf{N}_2 &= a_1\mathbf{A}^{n-1} \\ &\vdots \\ \mathbf{A}^2\mathbf{N}_{n-1} - \mathbf{A}\mathbf{N}_n &= a_{n-1}\mathbf{A} \\ \mathbf{A}\mathbf{N}_n &= a_n\mathbf{I}_n \end{aligned}$$

Adding up these terms shows that

$$0 = a_0\mathbf{A}^n + a_1\mathbf{A}^{n-1} + \cdots + a_{n-1}\mathbf{A} + a_n\mathbf{I} = p(\mathbf{A})$$

thereby completing our verification of the Cayley-Hamilton theorem.

The Cayley-Hamilton theorem is a useful technical tool. It implies for any $\mathbf{A} \in \mathbb{R}^{n\times n}$ that

$$\mathbf{A}^n = -a_1\mathbf{A}^{n-1} - a_2\mathbf{A}^{n-2} - \cdots - a_{n-1}\mathbf{A} - a_n\mathbf{I}_n$$

This can be used to show that

$$\mathbf{A}^{n+1} = (a_1^2 - a_2)\mathbf{A}^{n-1} + (a_1 a_2 - a_3)\mathbf{A}^{n-2} + \cdots + (a_1 a_{n-1} - a_n)\mathbf{A} + a_1 a_n \mathbf{I}_n$$

In other words, $\mathbf{A}^n$ and $\mathbf{A}^{n+1}$ can be expressed as a linear combination of a *finite* number of powers of $\mathbf{A}^k$ for $k = 0$ to $n-1$. So for any $k \geq n$, we can always write $\mathbf{A}^k$ as a linear combinations of $\mathbf{I}_n$, $\mathbf{A}$, $\mathbf{A}^2, \cdots, \mathbf{A}^{n-1}$. We will use this observation to obtain finite length representations of analytic functions, $f : \mathbb{C} \to \mathbb{C}$, of a complex variable.

A function of a complex variable, $f : \mathbb{C} \to \mathbb{C}$, is *analytic* if it has derivatives of all orders. If $f(s)$ is analytic then we can express it as a convergent power series of the form,

$$f(s) = \sum_{k=0}^{\infty} \alpha_k s^k, \quad \text{where } \alpha_k = \frac{1}{2\pi j} \int_{|w|=r} \frac{f(w)}{w^{k+1}} dw$$

This is, essentially, a Taylor series expansion of $f$ about $0$, which means that we can also write this as

$$f(s) = \sum_{k=0}^{\infty} \frac{f^{(k)}(s)}{k!} s^k$$

where $f^{(k)}(s) = \frac{d^k f(s)}{ds^k}$. We now show how the Cayley-Hamilton theorem can be used to reduce the order of an analytic function, $f(\mathbf{A})$, of the square matrix $\mathbf{A}$.

So consider the matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ and a *polynomial* $f : \mathbb{C} \to \mathbb{C}$ that takes values $f(s)$. Let $p(s) = \det(s\mathbf{I} - \mathbf{A})$ denote the characteristic polynomial of $\mathbf{A}$. Note that one can always write $f(s)$ in the form

$$f(s) = q(s)p(s) + r(s)$$

where $q(s)$ is found by polynomial long division, $p(s)$ is the characteristic polynomial, and the remainder polynomial, $r(s)$, is of degree less than or equal to $n-1$. The fact that such polynomials always exist is a consequence of the *division algorithm* in abstract algebra.

Now let $\lambda \in \mathbb{C}$ be any eigenvalue of $\mathbf{A}$. At such eigenvalues we know $p(\lambda) = 0$, and using this in the relation $f(s) = q(s)p(s) + r(s)$ we can conclude that

$$f(\lambda) = r(\lambda)$$

where $r(s)$ is the remainder polynomial of degree $n-1$ or less. Now consider the corresponding matrix polynomial $f(\mathbf{A})$ where one simply replaces $s$ by the matrix $\mathbf{A}$. From the Cayley-Hamilton theorem we know that $p(\mathbf{A}) = 0$ which means that

$$\begin{aligned} f(\mathbf{A}) &= q(\mathbf{A})p(\mathbf{A}) + r(\mathbf{A}) \\ &= r(\mathbf{A}) \end{aligned}$$

This last relationship can be used to set up a system of equations from which a reduced order representation of $f(\mathbf{A})$ can be obtained. Recall that $f$ is analytic, so the right hand side of the above equation is an infinite series formed from the Taylor series of $f$. By the division algorithm, on the other hand, we know that the polynomial on the left hand side of the above equation, $r$, has a degree less than or equal to $n-1$. This $r(\mathbf{A})$, therefore, represents our reduced order representation for the matrix function $f(\mathbf{A})$.

As an example, let us consider a polynomial function of $\mathbf{A}$

$$f(\mathbf{A}) = \mathbf{A}^4 + 3\mathbf{A}^3 + 2\mathbf{A}^2 + \mathbf{A} + \mathbf{I}$$

where the matrix $\mathbf{A} = \begin{bmatrix} 3 & 1 \\ 1 & 2 \end{bmatrix}$. The characteristic polynomial of $\mathbf{A}$ is

$$p(s) = \det(s\mathbf{I} - \mathbf{A}) = s^2 - 5s + 5$$

which means the eigenvalues of $\mathbf{A}$ are $\lambda_1 = 1.3820$ and $\lambda_2 = 3.6180$. We know that $f(s) = s^4 + 3s^3 + 2s^2 + 2s + 1$, so applying the division algorithm to find the remainder $r(s)$ gives

$$\begin{aligned} \frac{f(s)}{p(s)} &= \frac{s^4 + 3s^3 + 2s^2 + 2s + 1}{s^2 + 5s + 5} \\ &= s^2 + 8s + 37 + \frac{146s - 184}{s^2 - 5s + 5} = q(s) + \frac{r(s)}{p(s)} \end{aligned}$$

and so

$$f(s) = (s^2 + 8s + 37)p(s) + 146s - 184 = q(s)p(s) + r(s)$$

and so

$$r(s) = 146s - 184$$

Since we know for the given $\mathbf{A}$ that $f(\mathbf{A}) = r(\mathbf{A})$ we can conclude that

$$\begin{aligned} f(\mathbf{A}) &= \mathbf{A}^4 + 3\mathbf{A}^3 + 2\mathbf{A}^2 + \mathbf{A} + \mathbf{I} \\ &= 146\mathbf{A} - 184\mathbf{I} \end{aligned}$$

which is a reduced order representation of $f(\mathbf{A})$.

CHAPTER 2

# Linear Models for Dynamical Systems

**Linear systems theory** is concerned with linear mathematical models that are used to predict, simulate, and estimate the behavior of dynamical systems generating outputs as a function of *time*. Time may be either continuous (real-valued) or discrete (integer-valued). These systems are *dynamical* because their outputs at time $t$ are a function of the past outputs and past inputs, so these systems have "memory". In many cases, that "memory" can be encapsulated as a single real-valued vector, $x(t)$, that is sufficient to predict the system's outputs after time $t$. That vector, $x(t)$, is called the system's *state* at time $t$. The notion of "state" encapsulates the prior information needed to predict future outputs and it is one of the main concepts buried at the heart of linear systems theory. In many cases, we can implicitly characterize the state in terms of a differential or difference equation that forms the system's *state-space realization*. This chapter introduces state-based models for linear dynamical systems and discusses how they arise in the mathematical modeling of various systems found in engineering applications.

## 1. Linear State-based Realizations of Dynamical Systems

A dynamical system, $\mathbf{G}$, is one that accepts an input signal, $w$, and produces an output signal $y$. For us a *signal* will be a function of time, so that the input signal can be specified as $w : \mathbb{R} \to \mathbb{R}^m$ indicating that it maps time $t \in \mathbb{R}^m$ to a vector $w(t) \in \mathbb{R}^m$. We think of $w$ as the "name" of the signal and we think of $w(t)$ as the "value" that this signal takes at time $t$. The system, $\mathbf{G}$, therefore is an *operator* that transforms an input signal into an output

signal. In particular, $\mathbf{G}$ maps the input signal, $w$, onto an output signal $y$. We will use the notational convention $y = \mathbf{G}[w]$ to denote that the input signal $w$ was transformed into output signal $y$. The value that the output takes at time instant $t$ will be denoted as $y(t) = \mathbf{G}[w](t)$. We can make this more precise by specifying the set of all *input signals* as $\mathcal{L}_{\text{in}}$ and the set of output signals as $\mathcal{L}_{\text{out}}$. From this standpoint, a system may then be seen as a *signal transformation* and the system can be scoped out as $\mathbf{G} : \mathcal{L}_{\text{in}} \to \mathcal{L}_{\text{out}}$. The value that $\mathbf{G}$ takes for an input $w \in \mathcal{L}_{\text{in}}$ is denoted as $\mathbf{G}[w] \in \mathcal{L}_{\text{out}}$.

The preceding description of signals and systems is exceptionally abstract and provides no concrete way to "represent" a specific system. To quantitatively predict how a system might respond to a given input, we need a concrete way of mathematically representing the system. This course focuses on a particular kind of dynamic system, namely systems that are *linear* and systems that have *state space realizations*. A system is said to be *linear* if it satisfies the principle of superposition. In particular, let $\mathbf{G} : \mathcal{L}_{\text{in}} \to \mathcal{L}_{\text{out}}$ denote a system mapping signals from $\mathcal{L}_{\text{in}}$ onto signals in $\mathcal{L}_{\text{out}}$. Let us also assume that the signals in these two sets, $\mathcal{L}_{\text{in}}$ and $\mathcal{L}_{\text{out}}$, take values in the vector spaces $\mathbb{R}^m$ and $\mathbb{R}^p$, respectively. The system $\mathbf{G}$ is said to be linear if for any two signals $w_1, w_2 \in \mathcal{L}_{\text{in}}$ and any two real scalars, $\alpha, \beta \in \mathbb{R}$, we have

$$\mathbf{G}\left[\alpha w_1 + \beta w_2\right] = \alpha \mathbf{G}\left[w_1\right] + \beta \mathbf{G}\left[w_2\right] \tag{13}$$

In other words we can distribute the binary operations of scalar-vector multiplication and vector addition outside of the scope of the $\mathbf{G}$ operator. From an analysis standpoint this is extremely useful because it means that one can determine the response of the system to any input in terms of a linear combination of simpler inputs that whose behavior we already know.

We still, however, need a concrete way of representing the system itself. One way of taking care of this is through the system's *state-space realization*. The preceding system $\mathbf{G} : \mathcal{L}_{\text{in}} \to \mathcal{L}_{\text{out}}$ has a state space realization if the relation between the input and output signals can be said to satisfy the

following set of equations

$$\begin{aligned} \dot{x}(t) &= \mathbf{A}(t)x(t) + \mathbf{B}(t)w(t) \\ y(t) &= \mathbf{C}(t)x(t) + \mathbf{D}(t)w(t) \end{aligned} \tag{14}$$

for all $t \geq 0$. The signal $x : \mathbb{R} \to \mathbb{R}^n$ is an *internal signal* of the system called the system's *state* and $x(t)$, the state at time $t \in \mathbb{R}$ is a real-valued vector of dimension $n$ in $\mathbb{R}^n$. The other objects are matrix-valued functions of time, $\mathbf{A} : \mathbb{R} \to \mathbb{R}^{n\times n}$, $\mathbf{B} : \mathbb{R} \to \mathbb{R}^{n\times m}$, $\mathbf{C} : \mathbb{R} \to \mathbb{R}^{p\times n}$, and $\mathbf{D} : \mathbb{R} \to \mathbb{R}^{p\times m}$. For this system to be completely characterized we also need to specify the initial state at time 0, $x(0) = x_0$, and we need to specify the input signal $w$ for all time. But once this information is available then one can obtain a forward solution to the differential equation and use that to determine the output $y$ for all time greater than 0. Note that this state-space realization is characterized by the four matrix function we have given. So in particular, we can denote the state space realization of $\mathbf{G}$ (the operator) as a packed collection of matrices

$$\mathbf{G} \stackrel{s}{=} \left[\begin{array}{c|c} \mathbf{A}(t) & \mathbf{B}(t) \\ \hline \mathbf{C}(t) & \mathbf{D}(t) \end{array}\right]$$

We call this the packed matrix representation of linear system $\mathbf{G}$. The state space realization above has these system matrices varying over time, so this particular $\mathbf{G}$ is a time-varying system. If these matrices are constant for all time, then $\mathbf{G}$ is a time-invariant system.

**Remark:** The state-space realization in equation (14) assumed continuous-time inputs and outputs. State space realizations for discrete-time systems take a similar form.

$$\begin{aligned} x(k+1) &= \mathbf{A}(k)x(k) + \mathbf{B}(k)w(k) \\ y(k) &= \mathbf{C}(k)(k) + \mathbf{D}(k)w(k) \end{aligned} \tag{15}$$

The internal state $x : \mathbb{Z} \to \mathbb{R}^n$ is not a discrete-time function so that $x(k) \in \mathbb{R}^n$ denotes the system state at time instant $k \in \mathbb{Z}$.

Consider the state-space system $\mathbf{G} \stackrel{s}{=} \left[\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right]$. Is this system *linear*? To verify that this is indeed the case, we need to consider two arbitrary signals $w_1, w_2 \in \mathcal{L}_{\text{in}}$ and two scalars $\alpha, \beta \in \mathbb{R}$. For the inputs, $w_1$ and $w_2$, we know there are two state trajectories, $x_1$ and $x_2$, respectively, such that

$$\begin{aligned} \dot{x}_1(t) &= \mathbf{A}x_1(t) + \mathbf{B}w_1(t) \\ \dot{x}_2(t) &= \mathbf{A}x_2(t) + \mathbf{B}w_2(t) \\ y_1(t) &= \mathbf{C}x_1(t) + \mathbf{D}w_1(t) \\ y_2(t) &= \mathbf{C}x_2(t) + \mathbf{D}w_2(t) \end{aligned}$$

Now consider the signals

$$\begin{aligned} x(t) &= \alpha x_1(t) + \beta x_2(t) \\ w(t) &= \alpha w_1(t) + \beta w_2(t) \end{aligned}$$

then we have

$$\begin{aligned} \dot{x}(t) &= \alpha\dot{x}_1(t) + \beta\dot{x}_2(t) \\ &= \alpha\mathbf{A}x_1(t) + \alpha\mathbf{B}w_1(t) \\ &\quad \beta\mathbf{A}x_2(t) + \beta\mathbf{B}w_2(t) \\ &= \mathbf{A}(\alpha x_1(t) + \beta x_2(t)) + \mathbf{B}(\alpha w_1(t) + \beta w_2(t)) \\ &= \mathbf{A}x(t) + \mathbf{B}w(t) \end{aligned}$$

Moreover this means that the system's output under $w(t)$ will be

$$\begin{aligned} y(t) &= \mathbf{C}x(t) + \mathbf{D}w(t) \\ &= \mathbf{C}(\alpha x_1(t) + \beta x_2(t)) + \mathbf{D}(\alpha w_1(t) + \beta w_2(t)) \\ &= \alpha(\mathbf{C}x_1(t) + \mathbf{D}w_1(t)) + \beta(\mathbf{C}x_2(t) + \mathbf{D}w_2(t)) \\ &= \alpha y_1(t) + \beta y_2(t) \end{aligned}$$

The preceding two equations say that if we use $w(t) = \alpha w_1 + \beta w_2$ as an input to $\mathbf{G}$ that the output can be written as $y = \alpha y_1 + \beta y_2$ where $y_1$ is the response to $w_1$ and $y_2$ is the response to $w_2$. This assertion therefore verifies the principle of superposition in equation (13) and we can assert that any

state space realization $\left[\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right]$ is a realization for a linear system. This argument is summarized in the following theorem

THEOREM 1. *Let* $\mathbf{G} \stackrel{\text{s}}{=} \left[\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right]$ *is a system that satisfies the principle of superposition (13).*

The input and output signals of a dynamical system may be viewed as elements of a *linear space*. Linear spaces are abstract generalizations of the more familiar vector space concept. In our case, we will confine our attention to continuous-time systems whose inputs and outputs form linear spaces denoted as $\mathcal{L}(\mathbb{R}^m)$ and $\mathcal{L}(\mathbb{R}^p)$, respectively. The notation $\mathcal{L}(\mathbb{R}^n)$ being a linear space of *integrable functions*, whose elements are $x : \mathbb{R} \to \mathbb{R}^n$. Let us consider two *linear state-based* systems $\mathbf{G}_1 : \mathcal{L}(\mathbb{R}^m) \to \mathcal{L}(\mathbb{R}^p)$ and $\mathbf{G}_2 : \mathcal{L}(\mathbb{R}^m) \to \mathcal{L}(\mathbb{R}^p)$. We are going to define an *algebra* on such linear systems by introducing two different binary operations that we refer to as *parallel composition* and *cascade* or *series* composition.

The parallel composition of two systems, $\mathbf{G}_1$ and $\mathbf{G}_2$, will be denoted as "addition", with the symbol $\mathbf{G}_1 + \mathbf{G}_2$ and is defined on a componentwise manner with respect to the system's outputs. Namely, using the previous notational conventions representing systems as "signal transformations",

$$(\mathbf{G}_1 + \mathbf{G}_2)[w] = \mathbf{G}_1[w] + \mathbf{G}_2[w]$$

Note that on the left hand side of this equation, the symbol $+$ refers to the "addition" of the two systems whereas on the right hand side of the equation the $+$ symbol refers to addition over the linear space of output signals, $\mathcal{L}(\mathbb{R}^p)$. It is also convenient to think of this parallel composition in terms of a block diagram where each block represents a system transformation. Fig.1 shows the block diagram associated with $\mathbf{G}_1 + \mathbf{G}_2$. This diagram shows the two systems, $\mathbf{G}_1$ and $\mathbf{G}_2$, acting in parallel on the same input, $w$ before *adding* their outputs together.

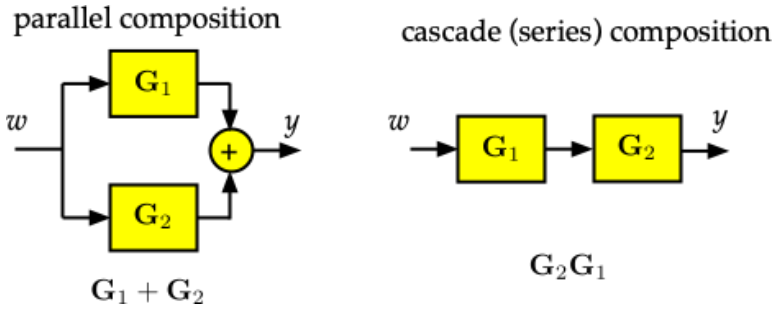



FIGURE 1. (left) parallel composition, $\mathbf{G}_1 + \mathbf{G}_2$: (right) series or cascade composition, $\mathbf{G}_2\mathbf{G}_1$

Since both of these systems are linear state-based systems, they each have a state space realization. So let us denote these two realizations as

$$\mathbf{G}_1 \stackrel{s}{=} \left[\begin{array}{c|c} \mathbf{A}_1 & \mathbf{B}_1 \\ \hline \mathbf{C}_1 & \mathbf{D}_1 \end{array}\right], \quad \mathbf{G}_2 \stackrel{s}{=} \left[\begin{array}{c|c} \mathbf{A}_2 & \mathbf{B}_2 \\ \hline \mathbf{C}_2 & \mathbf{D}_2 \end{array}\right]$$

Each of these realizations give rise to the following state based equations

$$\begin{aligned} \dot{x}_1 &= \mathbf{A}_1 x_1 + \mathbf{B}_1 w \\ \dot{x}_2 &= \mathbf{A}_2 x_2 + \mathbf{B}_2 w \\ y_1 &= \mathbf{C}_1 x_1 + \mathbf{D}_1 w \\ y_2 &= \mathbf{C}_2 x_2 + \mathbf{D}_2 w \end{aligned}$$

We know that $y_1 = \mathbf{G}_1[w]$ and $y_2 = \mathbf{G}_2[w]$. So the parallel composition of these two systems will produce the response

$$y = (\mathbf{G}_1 + \mathbf{G}_2)[w] = y_1 + y_2$$

in response to the input $w$. If we rewrite this output in terms of the state equations we get

$$\begin{aligned} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} &= \begin{bmatrix} \mathbf{A}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{bmatrix} w \\ y &= \begin{bmatrix} \mathbf{C}_1 & \mathbf{C}_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + (\mathbf{D}_1 + \mathbf{D}_2) w \end{aligned}$$

This means that we can immediately write down a state space realization for the parallel composition of two systems as

$$\mathbf{G}_1 + \mathbf{G}_2 \stackrel{\text{s}}{=} \left[\begin{array}{cc|c} \mathbf{A}_1 & \mathbf{0} & \mathbf{B}_1 \\ \mathbf{0} & \mathbf{A}_2 & \mathbf{B}_2 \\ \hline \mathbf{C}_1 & \mathbf{C}_2 & \mathbf{D}_1 + \mathbf{D}_2 \end{array}\right]$$

Note this is not the only state space realization we can obtain for the parallel decomposition, but it is a convenient one that can be readily programmed in a scripting language to automate the generation of a state-space realization for two systems connected in parallel.

The other binary operation is called a series or cascade composition of the systems $\mathbf{G}_1$ and $\mathbf{G}_2$. In this case we assume that $\mathbf{G}_1 : \mathcal{L}(\mathbb{R}^m) \rightarrow \mathcal{L}(\mathbb{R}^q)$ and $\mathbf{G}_2 : \mathcal{L}(\mathbb{R}^q) \rightarrow \mathcal{L}(\mathbb{R}^p)$. The series composition is denoted as the concatenation or multiplication of the two systems, denoted by the symbol $\mathbf{G}_2\mathbf{G}_1$. This binary operation is defined as

$$(\mathbf{G}_2\mathbf{G}_1)[w] = \mathbf{G}_2\left[\mathbf{G}_1\left[w\right]\right]$$

The block diagram in Fig. 1 shows that a cascade combination first has $\mathbf{G}_1$ transform the input $w$ and that $\mathbf{G}_2$ acts to transform the output of $\mathbf{G}_1$.

Since both of these systems are linear state-based systems, we can use their state space realizations to construct a state-space realization of the cascaded system $\mathbf{G}_2$. These realizations give rise to the following state-based equations

$$\begin{aligned} \dot{x}_1 &= \mathbf{A}_1 x_1 + \mathbf{B}_1 w \\ \dot{x}_2 &= \mathbf{A}_2 x_2 + \mathbf{B}_2 y_1 \\ y_1 &= \mathbf{C}_1 x_1 + \mathbf{D}_1 w \\ y &= \mathbf{C}_2 x_2 + \mathbf{D}_2 y_1 \end{aligned}$$

Note that we have modified the inputs to the second and fourth equations to be $y_1$, the output from the first system, $\mathbf{G}_1$. What we see here is that $y_1$ is an intermediate variable that we would like to remove from the representation.

So substituting the third equation into the second and fourth yields,

$$\begin{aligned}
\dot{x}_1 &= \mathbf{A}_1 x_1 + \mathbf{B}_1 w \\
\dot{x}_2 &= \mathbf{A}_2 x_2 + \mathbf{B}_2 (\mathbf{C}_1 x_1 + \mathbf{D}_1 w) \\
&= \mathbf{B}_2\mathbf{C}_1 x_1 + \mathbf{A}_2 x_2 + \mathbf{B}_2\mathbf{D}_1 w \\
y &= \mathbf{C}_2 x_2 + \mathbf{D}_2 (\mathbf{C}_1 x_1 + \mathbf{D}_1 w) \\
&= \mathbf{D}_2\mathbf{C}_1 x_1 + \mathbf{C}_2 x_2 + \mathbf{D}_2\mathbf{D}_1 w
\end{aligned}$$

These equations immediately give rise to the following state space realization for the cascaded system

$$\mathbf{G}_2\mathbf{G}_1 \stackrel{\text{s}}{=} \left[\begin{array}{cc|c} \mathbf{A}_1 & \mathbf{0} & \mathbf{B}_1 \\ \mathbf{B}_2\mathbf{C}_1 & \mathbf{A}_2 & \mathbf{B}_2\mathbf{D}_1 \\ \hline \mathbf{D}_2\mathbf{C}_1 & \mathbf{C}_2 & \mathbf{D}_2\mathbf{D}_1 \end{array}\right]$$

These two binary operations can be used to construct a wide range of systems. The fact that we have two "formulae" for constructing state space realizations of these system compositions means we can automate the construction of state space realizations for any system that can be represented as a sequence of parallel or cascade compositions. In other words, these results provide a useful algorithmic way for constructing state space realizations. For convenience we will summarize the preceding derivation as a theorem.

THEOREM 2. *Let* $\mathbf{G}_1 \stackrel{\text{s}}{=} \left[\begin{array}{c|c} \mathbf{A}_1 & \mathbf{B}_1 \\ \hline \mathbf{C}_1 & \mathbf{D}_1 \end{array}\right]$ *and* $\mathbf{G}_2 \stackrel{\text{s}}{=} \left[\begin{array}{c|c} \mathbf{A}_2 & \mathbf{B}_2 \\ \hline \mathbf{C}_2 & \mathbf{D}_2 \end{array}\right]$*, then state space realizations for the parallel and cascade composition of these two systems (assuming the input/outputs of each subsystem are correctly dimensioned) are*

$$\mathbf{G}_1 + \mathbf{G}_2 \stackrel{\text{s}}{=} \left[\begin{array}{cc|c} \mathbf{A}_1 & \mathbf{0} & \mathbf{B}_1 \\ \mathbf{0} & \mathbf{A}_2 & \mathbf{B}_2 \\ \hline \mathbf{C}_1 & \mathbf{C}_2 & \mathbf{D}_1 + \mathbf{D}_2 \end{array}\right], \qquad \mathbf{G}_2\mathbf{G}_1 \stackrel{\text{s}}{=} \left[\begin{array}{cc|c} \mathbf{A}_1 & \mathbf{0} & \mathbf{B}_1 \\ \mathbf{B}_2\mathbf{C}_1 & \mathbf{A}_2 & \mathbf{B}_2\mathbf{D}_1 \\ \hline \mathbf{D}_2\mathbf{C}_1 & \mathbf{C}_2 & \mathbf{D}_2\mathbf{D}_1 \end{array}\right]$$

## 2. Transform Modeling of Time-invariant Linear Systems

Let us consider a time-invariant system $\mathbf{G} \stackrel{s}{=} \left[\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right]$ whose state equations take the form

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

We first want to predict how the system will behave in response to a known applied input signal $w$. How does one go about solving the ordinary differential equation (ODE)? This section reviews the use of Laplace transform methods in solving the state equations for continuous time systems.

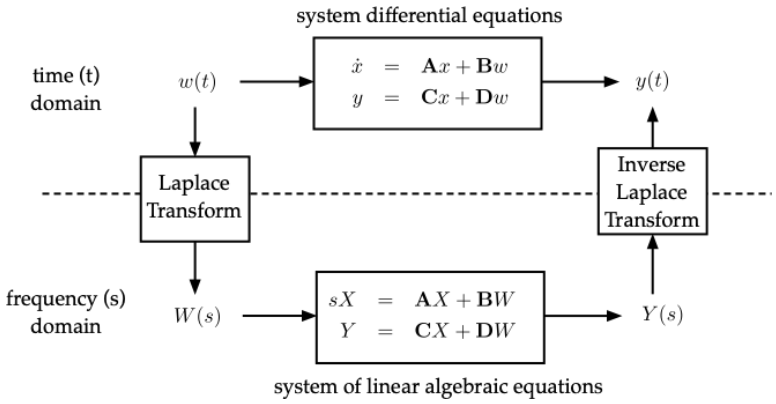



FIGURE 2. Transform Method for Solving Ordinary Differential Equations

Most undergraduate students who have taken an elementary course in ordinary differential equations should be familiar with the use of transform methods to solve systems of constant-coefficient ordinary differential equations. Fig. 2 is a commutative diagram illustrating how the transform method works. The diagram shows two ways for obtaining the system's response $y(t)$, to an applied input $w(t)$. The first way directly solves the differential equation in the time domain and therefore requires calculus to get a solution. The second approach transforms the time-domain signal $w$

through a single-sided Laplace transform into a function of a complex variable, $W$. It transforms the differential equations into a system of linear algebraic equations (LAE). This LAE system can be solved more easily than the original differential equations to obtain a function $Y$ that is the Laplace transform of the output signal $y$. One would only need to inverse transform $Y$ back into $y$ to complete the problem. There are two things which make this approach algorithmically more attractive than directly solving the ODEs. First, the transformed system differential equations form a set of linear algebraic equations that can be easily solved using methods from high school mathematics. Second the linear nature of the system allows us to easily compute the Laplace and inverse Laplace transforms of the signals. Together these two facts provide a powerful set of tools that solve the system state equations and also provide significant insight into how one might "control" that system to force it to behave in a desired manner.

Transform methods for continuous-time systems are based on the single sided Laplace transform while discrete-time systems may be modeled using single sided $z$ transforms. The following subsections provides an informal review of single sided Laplace transforms and $z$-transform methods.

**2.1. Single-sided Laplace Transforms.** A single sided Laplace transform is an invertible linear transformation

$$\mathcal{L} \,:\, L(\mathbb{R}) \to L(\mathbb{C})$$

where $L(\mathbb{R})$ is a linear space of integrable time domain signals and $L(\mathbb{C})$ represents the set of functions of a complex variables with removable singularities. We will refer to the real-valued function, $g : \mathbb{R} \to \mathbb{R}$ as a *time-domain* function (signal) and we will refer to its image under the Laplace transform, $\mathcal{L}[g] \,:\, \mathbb{C} \to \mathbb{C}$, as a *frequency-domain function*. The single sided Laplace transform of the function $g$ takes the following values

$$\mathcal{L}[g](s) = \int_0^\infty g(t)e^{-st}dt$$

where $s \in \mathbb{C}$.

Let us consider the signal $y(t) = e^{-\alpha t}u(t)$ where $u$ is a *unit step function* that takes values

$$u(t) = \begin{cases} 1 & t \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

The signal $y(t)$ is zero for $t < 0$, jumps to 1 when $t = 0$, and then decays at an exponential rate for $t > 0$ if $\alpha > 0$ and grows at an exponential rate if $\alpha < 0$. This means that the integral equation only exists when $\alpha > 0$. In that case we have

$$\begin{aligned} Y(s) = \mathcal{L}[y](s) &= \int_0^\infty e^{-(s+\alpha)t}dt \\ &= \int_0^\infty e^{-(\sigma+\alpha)t}e^{-j\omega t}dt \\ &= \int_0^\infty e^{-(\sigma+\alpha)t}\cos(\omega t)dt - j\int_0^\infty e^{-(\sigma+\alpha)t}\sin(\omega t)dt \end{aligned}$$

where we used the variable substitution $s = \sigma + j\omega$ with $\sigma, \omega \in \mathbb{R}$ and we used the Euler relation $e^{-j\omega t} = \cos(\omega t) - j\sin(\omega t)$. Note that the integrals in the last equation are standard Riemann integrals that exist if and only if $\sigma + \alpha > 0$. The set of complex values, $s = \sigma + j\omega$, that satisfy this relation form the *region of convergence* (ROC) of the transform.

$$\text{RoC} = \{s = \sigma + j\omega \,:\, \sigma > -\alpha\}$$

In particular, this means that the value $Y(s)$ only exists (is finite) if $s$ lies in the RoC. If $s$ is chosen outside of RoC, then $Y(s)$ is not well defined. The value that this integral converges to can be readily shown to be

$$\begin{aligned} \mathcal{L}\left[e^{-\alpha t}u(t)\right](s) &= \int_0^\infty e^{-(s+\alpha)t}dt = -\frac{1}{s+\alpha}e^{-(s+\alpha)t}\Big|_0^\infty \\ &= -\frac{1}{\sigma+\alpha+j\omega}\, e^{-(\sigma+\alpha)t}(\cos(\omega t) - j\sin(\omega t))\Big|_0^\infty \\ &= \frac{1}{s+\alpha}, \quad \text{for } \text{Re}(s) > -\alpha \end{aligned}$$

Determining the Laplace transform of a given function using the preceding integral formula is cumbersome and does not need to be repeated for

| type | $x(t)$ | $F(s)$ |
|---|---|---|
| impulse | $\delta(t)$ | $1$ |
| step | $u(t)$ | $\frac{1}{s}$ |
| ramp | $tu(t)$ | $\frac{1}{s^2}$ |
| exponential | $e^{-at}u(t)$ | $\frac{1}{s+a}$ |
| sine | $\sin(\omega t)u(t)$ | $\frac{\omega}{s^2+\omega^2}$ |
| cosine | $\cos(\omega t)u(t)$ | $\frac{s}{s^2+\omega^2}$ |
| damped ramp | $te^{-at}u(t)$ | $\frac{1}{(s+a)^2}$ |
| damped general ramp | $t^n e^{-at}u(t)$ | $\frac{n!}{(s+a)^{n+1}}$ |
| damped sine | $e^{-at}\sin(\omega t)u(t)$ | $\frac{\omega}{(s+a)^2+\omega^2}$ |
| damped cosine | $e^{-at}\cos(\omega t)u(t)$ | $\frac{(s+a)}{(s+a)^2+\omega^2}$ |

TABLE 1. Table of Standard Single Sided Laplace Transforms

functions that are closely related to $e^{-\alpha t}u(t)$. Rather than directly integrating any given function, we formally compute the integrals for a *canonical* set of functions and then use a set of *operational transforms* to determine the transform of the given function. Table 1 shows a table of canonical single-sided Laplace transform pairs. Table 2 shows at table of well known *operational transform pairs*.

The tables 1 and 2 are used together to compute more complex transform pairs that may not be in the original table. As an example, let us consider the function

$$y(t) = e^{-3t}\cos(5t + 30^\circ)u(t)$$

The first thing we note is that

$$\cos(\omega t + \phi) = \cos(\phi)\cos(\omega t) - \sin(\phi)\sin(\omega t)$$

we let $\omega = 5$ and $\phi = 30^\circ$ so this becomes

$$\cos(5t + 30^\circ) = \frac{\sqrt{3}}{2}\cos(5t) - \frac{1}{2}\sin(5t)$$

and we can rewrite $y(t)$ as

$$y(t) = \frac{\sqrt{3}}{2}e^{-3t}\cos(5t)u(t) - \frac{1}{2}e^{-3t}\sin(5t)u(t)$$

From the table of operational transforms we know that

$$Y(s) = \frac{\sqrt{3}}{2}\mathcal{L}\left[e^{-3t}\cos(5t)u(t)\right] - \frac{1}{2}\mathcal{L}\left[e^{-3t}\sin(5t)u(t)\right]$$

From the table of canonical transforms we then get

$$Y(s) = \frac{\sqrt{3}}{2}\frac{s+3}{(s+3)^2+25} - \frac{1}{2}\frac{5}{(s+3)^2+25}$$

When $Y(s)$ is a rational function (the ratio of two polynomials in $s$), it is customary practice to simply the expression into a single ratio of two monic polynomials multiplied by a real constant. So using this convention $Y(s)$ becomes

$$Y(s) = \frac{\sqrt{3}}{2}\frac{s+5.8868}{s^2+6s+34}$$

and the region of convergence includes that $s$ where $\mathrm{Re}(s) > -3$.

An important aspect of the Laplace transform is that it is invertible. This means that there exists a linear transformation $\mathcal{L}^{-1} : L(\mathbb{C}) \to L(\mathbb{R})$ such that $\mathcal{L}^{-1}[Y] = y$ if and only if $\mathcal{L}[y] = Y$. This inverse transform is also characterized by an integral formula,

$$y(t) = \mathcal{L}^{-1}[Y](t) = \frac{1}{2\pi}\int_{\sigma-j\infty}^{\sigma+j\infty} Y(s)e^{st}ds$$

where $\sigma = \mathrm{Re}(s)$ is for any $s$ in the RoC of $Y(s)$. One may readily verify this formula when $Y(s)$ is *strictly proper*; namely $|Y(s)| \to 0$ as $|s| \to \infty$. To verify this integral equation, we create a *contour*, $C_R$, that is contained within the transform's region of convergence. One can show that the ROC for a strictly proper Laplace transform is a half-space in the complex plane, so we construct our contour from two parts. The first part is the line $s = \sigma + j\omega$ where $\sigma$ is the real part of any complex number in the ROC and

| Type | $h(t)$ | $H(s)$ |
|---|---|---|
| Linearity | $\alpha f(t) + \beta g(t)$ | $\alpha F(s) + \beta G(s)$ |
| Time Shift | $f(t-T)$ | $e^{-sT}F(s)$ |
| | for $T > 0$ | |
| $s$-shifting | $e^{-at}f(t)u(t)$ | $F(s+a)$ |
| Time scale | $f(at), \quad a > 0$ | $\frac{1}{a}F\left(\frac{s}{a}\right)$ |
| Differentiation | $\frac{df(t)}{dt}$ | $sF(s) - f(0^-)$ |
| Integration | $\int_0^t f(\tau)d\tau$ | $\frac{F(s)}{s}$ |
| Initial Value | $\lim_{t\to 0^+} f(t)$ | $\lim_{s\to\infty} sF(s)$ |
| | assuming no impulse functions | |
| Final Value | $\lim_{t\to\infty} f(t)$ | $\lim_{s\to 0} sF(s)$ |
| | | assuming at most one pole at the origin |
| | | with remaining poles having negative real parts |
| Convolution | $\int_0^t x(t-\tau)g(\tau)d\tau$ | $F(s)G(s)$ |

TABLE 2. Table of Operational Transforms (Laplace)

where $\omega \in [-R, R]$ where $R$ is large. To close the contour we take a large half circle, $C_R$, of radius $R$ that connects the points $\sigma + jR$ and $\sigma - jR$. We can then invoke the Cauchy integral theorem [1] to assert

$$\begin{aligned} Y(s) &= \lim_{R\to\infty} \frac{1}{2\pi j} \oint_{C_R} \frac{F(w)}{w-s} dw \\ &= \lim_{R\to\infty} \frac{1}{2\pi j} \int_{C_R} \frac{F(w)}{s-w} + \lim_{R\to infty} \frac{1}{2\pi j} \int_{\sigma-jR}^{\sigma+jR} \frac{F(w)}{s-w} dw \end{aligned}$$

[1] Cauchy Integral theorem [Levinson and Redheffer (1970)]: If $f(z)$ is an analytic function of a complex variable in a simply connected domain, thenf or any simple closed contour in that domain, the contour integral of $f$ is zero.

One can show that the first integral goes to zero as $R \to \infty$ because $Y(s)$ is strictly proper, which leaves

$$Y(s) = \frac{1}{2\pi j} \int_{\sigma - j\omega}^{\sigma + j\infty} \frac{F(w)}{s - w} dw$$

We now insert this into the following equation

$$\begin{aligned} f(t) = \mathcal{L}^{-1}\left[\mathcal{L}[y]\right](t) &= \mathcal{L}^{-1}\left[\frac{1}{2\pi j} \int_{\sigma - j\omega}^{\sigma + j\omega} \frac{Y(w)}{s - w} dw\right] \\ &= \frac{1}{2\pi j} \int_{\sigma - j\omega}^{\sigma + j\omega} Y(w) \mathcal{L}^{-1}\left[\frac{1}{s - w}\right] dw \\ &= \frac{1}{2\pi j} \int_{\sigma - j\omega}^{\sigma + j\omega} Y(w) e^{wt} dw \end{aligned}$$

which verifies the integral formula we gave for the inverse Laplace transform.

We can now use the transform pairs in tables 1 and 2 to solve for the state trajectory $x(t)$ of a linear dynamical system. As an example let us consider the following state-based realization

$$\begin{aligned} \begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} &= \begin{bmatrix} 4 & -5 \\ 2 & -3 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u(t), \quad x_0 = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \\ y &= \begin{bmatrix} 1/2 & 1 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} \end{aligned}$$

where $u(t)$ is a unit step function. A common way of solving such differential equations is to use the transform method illustrated in Fig. 2. This approach uses the differentiation operational transform in table 2 to transform the differential equation into a system of linear algebraic equations

$$\begin{aligned} sX_1(s) - 2 &= 4X_1(s) - 5X_2(s) + \frac{1}{s} \\ sX_2(s) - 1 &= 2X_1(s) - 3X_2(s) \\ Y(s) &= \frac{1}{2}X_1(s) + X_2(s) \end{aligned}$$

The first two equations for a system of linear algebraic equations that we solve for $X_1(s)$ and $X_2(s)$ using standard methods such as Gaussian elimination. In particular, these two equations can be rewritten as

$$\begin{bmatrix} \frac{1}{s}+2 \\ 1 \end{bmatrix} = \begin{bmatrix} s-4 & 5 \\ -2 & s+3 \end{bmatrix} \begin{bmatrix} X_1(s) \\ X_2(s) \end{bmatrix}$$

Since the matrix on the right hand side is 2 by 2, we can use Cramer's formula[2] to get its inverse

$$\begin{aligned} \begin{bmatrix} X_1(s) \\ X_2(s) \end{bmatrix} &= \begin{bmatrix} s-4 & 5 \\ -2 & s+3 \end{bmatrix}^{-1} \begin{bmatrix} \frac{1+2s}{s} \\ 1 \end{bmatrix} \\ &= \frac{1}{(s-4)(s+3)+10} \begin{bmatrix} s+3 & -5 \\ 2 & s-4 \end{bmatrix} \begin{bmatrix} 2\frac{s+1/2}{s} \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} 2\frac{s^2+s+1.5}{s(s-2)(s+1)} \\ \frac{s^2+2}{s(s-2)(s+1)} \end{bmatrix} \end{aligned}$$

**Remark:** We can use software tools such as MATLAB to do much of the "algebra" used to obtain the preceding example's final expression. In particular, note the original system equations can be written in matrix-vector form as

$$\begin{aligned} \dot{x} &= \mathbf{A}x + \mathbf{B}u \\ y &= \mathbf{C}x \end{aligned}$$

Taking the Laplace transform gives

$$\begin{aligned} sX(s) - x_0 &= \mathbf{A}X(s) + \mathbf{B}\frac{1}{s} \\ Y(s) &= \mathbf{C}X(s) \end{aligned}$$

where $x_0 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$, $\mathbf{A} = \begin{bmatrix} 4 & -5 \\ 2 & -3 \end{bmatrix}$, $\mathbf{B} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\mathbf{C} = \begin{bmatrix} 1/2 & 1 \end{bmatrix}$. So we can use the following MATLAB script

[2] If $\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, then $\mathbf{A}^{-1} = \frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$

```
s = tf('s');
A = [4 -5; 2 -3];
B = [1;0];
x0 = [2;1];
X = inv(s*eye(2,2)-A)*(B*1/s+x0);
zpk(X)
```

The last command converts the expression for `X` into a zero-pole form that factors the numerator and denominator polynomials

```
ans =

  From input to output...
       2 (s-2) (s+1) (s^2 + s + 1.5)
   1:  -----------------------------
             s (s-2)^2 (s+1)^2

       (s-2) (s+1) (s^2 + 2)
   2:  ---------------------
         s (s-2)^2 (s+1)^2

Continuous-time zero/pole/gain model.
```

Note that the computer algebra done by MATLAB does not cancel like terms between the numerator and denominator, so this must be done by hand to get

$$X_1(s) = 2\frac{s^2+s+1.5}{s(s-2)(s+1)}, \quad X_2(s) = \frac{s^2+2}{s(s-2)(s+1)}$$

Since $Y(s) = X_1(s)/2 + X_2(s)$, we can readily conclude that the system response to the step input is

$$Y(s) = \frac{s^2+s+1.5}{s(s-2)(s+1)} + \frac{s^2+2}{s(s-2)(s+1)} = 1.5\frac{s^2+0.333s+1.833}{s(s-2)(s+1)}$$

Again, I used MATLAB and the `zpk` command to get the final expression for $Y(s)$.

Because the Laplace transform is invertible, we can readily invert the transform in our example to obtain a time-domain representation of the preceding transform. For cases where the transform is a strictly proper rational transform, we can readily use the partial fraction expansion (PFE) method to rewrite our rational expression as a sum of elementary transforms that are in Table 1. This means that we write our earlier expression for $Y(s)$ as

$$\begin{aligned} Y(s) &= 1.5\frac{s^2+0.333s+1.833}{s(s-2)(s+1)} \\ &= \frac{K_0}{s}+\frac{K_1}{s-2}+\frac{K_2}{(s+1)} \end{aligned}$$

where $K_0$, $K_1$, and $K_2$, are real-valued coefficients. A standard way of teaching undergraduate students how to evaluate these coefficients is to rewrite our preceding expression as

$$\begin{aligned} Y(s) &= \frac{K_0}{s}+\frac{K_1}{s-2}+\frac{K_2}{s+1} \\ &= \frac{K_0(s-2)(s+1)+K_1s(s+1)+K_2s(s-2)}{s(s-2)(s+1)} \\ &= \frac{s^2(K_0+K+1+K_2)+s(-K_0+K_1-2K_2)+(-2K_0)}{(s(s-2)(s+1)} \\ &= 1.5\frac{s^2+0.333s+1.833}{s(s-2)(s+1)} \end{aligned}$$

we would then equate coefficients of the numerator polynomial, set up a system of linear algebraic equations in terms of the coefficients and then solve for those coefficients.

Another approach for finding these coefficients uses the Residue calculus[3] from complex analysis. The residue calculus states that the coefficients,

[3] In complex analysis, the Residue theorem [Levinson and Redheffer (1970)] is a tool used to evaluate line integrals of analytic functions over closed curves.

$K_0$, $K_1$, and $K_2$ are the residues of the rational function, where

$$\begin{aligned} K_0 &= \lim_{s\to 0} 1.5\frac{s^2+0.333s+1.833}{(s-2)(s+1)} \\ &= \frac{1.5(1.833)}{-2} = -1.3750 \\ K_1 &= \lim_{s\to 2} 1.5\frac{s^2+0.333s+1.833}{s(s+1)} \\ &= 1.5\frac{4+.6666+1.8333}{2(3)} = 1.6250 \\ K_2 &= \lim_{s\to -1} 1.5\frac{s^2+0.3333s+1.8333}{s(s-2)} \\ &= 1.5\frac{1-.333s+1.8333}{-1(-3)} = 1.25 \end{aligned}$$

This tends to be easier to solve by hand and it allows us to immediately see that

$$Y(s) = \frac{-1.3750}{s} + \frac{1.6250}{s-2} + \frac{1.25}{s+1}$$

Each term on the right hand side is in the Laplace transform table, so we can immediately write this out as

$$y(t) = -1.3750u(t) + 1.6250e^{2t}u(t) + 1.25e^{-t}u(t)$$

What we just showed is that any time-invariant state-based system of the form

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

can be solved using the transform method. The Laplace transform of the state equations gives

$$\begin{aligned} sX(s) - x(0) &= \mathbf{A}X(s) + \mathbf{B}W(s) \\ Y(s) &= \mathbf{C}X(s) + \mathbf{D}W(s) \end{aligned}$$

These equations are algebraic and we solve the first one to show

$$\begin{aligned} X(s) &= (s\mathbf{I}-\mathbf{A})^{-1}(\mathbf{B}W(s)+x(0)) \\ &= (s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}W(s)+(s\mathbf{I}-\mathbf{A})^{-1}x(0) \end{aligned}$$

inserting our expression for $X(s)$ into the second state equation gives

$$\begin{aligned} Y(s) &= \mathbf{C}(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}W(s)+\mathbf{C}(s\mathbf{I}-\mathbf{A})^{-1}x(0) \\ &= \mathbf{G}(s)W(s)+\mathbf{G}_0(s)x(0) \end{aligned}$$

The first term $\mathbf{G}(s)$ is called the transfer function of the system from the input $w$ to the output. It represents the system's *zero-state* or *forced* response to the external input $w$. The second term, $\mathbf{G}_0$ characterizes the system's *zero-input* or *natural* response to the system's initial states. It may also be thought of as a transfer function from the input $x_0$ to the output. The transfer function $\mathbf{G}(s)$ is usually seen as a *concrete representation* of the system's input-output behavior (assuming zero initial condition). We summarize this in the following result

THEOREM 3. *Given the state-space realization* $\mathbf{G} \stackrel{s}{=} \left[\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right]$, *then the transfer function of this system is*

$$\mathbf{G}(s)=\mathbf{C}(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}+\mathbf{D}$$

The transfer function from $W(s)$ to $Y(s)$ has an explicit form in our example which we can readily derive

$$\begin{aligned} \mathbf{G}(s) &= \mathbf{C}(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}+\mathbf{D} \\ &= \begin{bmatrix} 1/2 & 1 \end{bmatrix}\begin{bmatrix} s-4 & 5 \\ -2 & s+3 \end{bmatrix}^{-1}\begin{bmatrix} 1 \\ 0 \end{bmatrix} \\ &= \frac{1}{(s-4)(s+3)+10}\begin{bmatrix} 1/2 & 1 \end{bmatrix}\begin{bmatrix} s+3 & -5 \\ 2 & s-4 \end{bmatrix}\begin{bmatrix} 1 \\ 0 \end{bmatrix} \\ &= \frac{0.5s+3.5}{s^2-s-2}=0.5\frac{s+7}{(s-2)(s+1)} \end{aligned}$$

The response of this system to a given input $w(t)$ assuming zero initial conditions $x(0) = 0$ will be the inverse transform of

$$Y(s) = \mathbf{G}(s)W(s)$$

This is in our table of operational transforms which implies

$$y(t) = \int_0^t w(t-\tau)g(\tau)d\tau$$

where $g$ is the the inverse transform of $\mathbf{G}(s)$. Of greater importance is the fact that the above integral can be viewed as a binary operation on $w(t)$ and $g(t)$ called the convolution. The inverse transform $g$ is called the *impulse response function* of the system because if we let $w(t) = \delta(t)$, then $Y(s) = G(s)$ and we see that $g(t)$ becomes the response of the system to a unit impulse input, $\delta(t)$.

**Remark:** We could have also defined a linear system in terms of the convolution of the input with the system's impulse response function. In general, this would take the form

$$y(t) = \int_{-\infty}^t g(t,\tau)w(\tau)d\tau$$

to imply that the impulse response function may also be a function of the time $t$, by itself. This would imply that the "response" changes depending upon when the impulse is applied and would mean that the system is *time-varying*. In particular, we can readily see that if $g$ is only a function of the time difference $t - \tau$ in the integral then the system is *time-invariant*.

**2.2. Single-Sided $z$-Transforms:** Laplace transforms are used for continuous-time signals. A similar transform-based tool exists for discrete-time signal, $y : \mathbb{Z} \to \mathbb{R}$. The tool we'll use below is called the $z$-transform. The single sided $z$-transform is given by

$$\mathcal{Z}[y](z) = \sum_{k=0}^{\infty} y(k)z^{-k}$$

| type | $y(t)$ | $Y(s)$ |
|---|---|---|
| impulse | $\delta(k)$ | $1$ |
| step | $u(k)$ | $\frac{z}{z-1}$ |
| geometric | $\alpha^k u(k)$ | $\frac{z}{z-\alpha}$ |
| damped ramp | $k\alpha^k u(k)$ | $\frac{\alpha z}{(z-\alpha)^2}$ |
| cosine | $\cos(\omega k)u(k)$ | $\frac{z^2-\cos(\omega)z}{z^2-2\cos(\omega)z+1}$ |
| sine | $\sin(\omega k)u(k)$ | $\frac{\sin(\omega)z}{z^2-2\cos(\omega)z+1}$ |
| damped cosine | $\alpha^k\cos(\omega k)u(k)$ | $\frac{z^2-\alpha\cos(\omega)z}{z^2-2\alpha\cos(\omega)z+\alpha^2}$ |
| damped cosine | $\alpha^k\sin(\omega k)u(k)$ | $\frac{\alpha\sin(\omega)z}{z^2-2\alpha\cos(\omega)z+\alpha^2}$ |

TABLE 3. Canonical $z$-transform pairs

As an example we compute the $z$-transform for the signal

$$y(k) = \alpha^k u(k)$$

where $u$ is a unit step function. If $|\alpha| < 1$ then $|y|$ asymptotically approaches zero as $k \to \infty$. If $|\alpha| > 1$, then $|y|$ increases in a geometric manner as $k \to \infty$. We want to compute the single sided $z$-transform of this function

$$Y(z) = \sum_{k=0}^{\infty} \alpha^k z^{-k} = \sum_{k=0}^{\infty} (\alpha z^{-1})^k$$

The series is convergent when there exists $M$ such that $\sum_{k=0}^{\infty} |\alpha z^{-1}|^k = M < \infty$. This occurs when $|\alpha z^{-1}| < 1$ or equivalently $|z| > |\alpha|$. The value that the series converges to may be computed as follows.

$$Y(z) = 1 + \alpha z^{-1} + (\alpha z^{-1})^2 + (\alpha z^{-1})^3 + \cdots$$

Notice that

$$\alpha z^{-1} Y(z) = \alpha z^{-1} + (\alpha z^{-1})^2 + (\alpha z^{-1})^3 + \cdots$$

If we subtract the last two equations we get

$$Y(z) - \alpha z^{-1} Y(z) = 1$$

and solving for $Y(z)$ gives

$$Y(z) = \sum_{k=0}^{\infty} (\alpha z^{-1})^k = \frac{1}{1 - \alpha z^{-1}}$$

We sometimes prefer expressing $Y(z)$ as a rational function in terms of the indeterminate variable, $z$, which gives

$$Y(z) = \frac{z}{z - \alpha}$$

Determining $z$-transforms using the series equation is also somewhat cumbersome. So the usual approach is to determine transforms for a select set of canonical signals and use these canonical transforms along with a corresponding set of operational transforms to find the transform for more complex functions. A basic table of $z$-transform pairs is given in table 3. A basic table of operational $z$-transform pairs is given in table 4.

Table 3 provides a set of elementary transform pairs that can be seen as a basic dictionary of relationships. Not all signals may be in this form, though they may be closely related. There are, therefore, a number of **operational transforms** that can be used to find the transform of a function that is "related" to that of the function in the table. A table of these operational transforms is provided in table 4. What should be apparent is that the basic methods used by the $z$-transform are similar to those of the Laplace transform.

The preceding example confined its attention to continuous-time systems, but all of the prior discussion is also relevant to discrete-time systems. For discrete-time systems, however, we use the $z$-transform. So let us consider the discrete-time system whose input/output signals satisfy the equation

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

| Type | $h(k)$ | $H(z)$ |
|---|---|---|
| Linearity | $\alpha f(k) + \beta g(k)$ | $\alpha F(z) + \beta G(z)$ |
| Time Delay | $f(k-1)$ | $z^{-1}F(z) + f(-1)$ |
| Time Advance | $f(k+1)$ | $zF(z) - zf(0)$ |
| $z$-scaling | $e^{j\omega k} f(k)$ | $F(e^{-j\omega} z)$ |
| First difference | $f(k) - f(k-1)$ | $(1 - z^{-1})F(z) - f(-1)$ |
| Accumulation | $\sum_{k=0}^{n} f(k)$ | $\frac{1}{1-z^{-1}} F(z)$ |
| ramped function | $kx(k)$ | $-z\frac{dF(z)}{dz}$ |
| Convolution Sum | $y(k) = \sum_{n=0}^{k} f(n)g(k-n)$ | $F(z)G(z)$ |
| Initial Value | $f(0)$ | $\lim_{z\to\infty} F(z)$ |
| Final Value | $\lim_{k\to\infty} f(k)$ | $\lim_{z\to 1}(z-1)F(z)$ |

TABLE 4. Table of Operational Transforms ($z$)

where $k \geq 0$ and where $y(0) = 0$ and $w$ is an input function of the form

$$w(k) = \begin{cases} 1 & \text{for } k = 1, 2, \ldots \\ 0 & \text{for } k \leq 0 \end{cases}$$

Note that this is a unit step function that has been shifted so it starts at $k = 1$, rather than $k = 0$. We denote the $z$-transforms of $y$ and $w$ as $\hat{y}$ and $\hat{w}$, respectively. We can use the operational transforms in table 4 to see that

$$z\hat{y}(z) \quad = \quad -\hat{y}(z)\frac{1}{3} + \hat{w}(z)$$

Solving for $\hat{y}(z)$ gives

$$\hat{y}(z)\left(z + \frac{1}{3}\right) = \hat{w}(z)$$

which we rewrite as

$$\hat{y}(z) = \frac{1}{z + 1/3}\hat{w}(z)$$

from which we see the system's transfer function is $\mathbf{G}(z) = \frac{1}{z+1/3}$.

Since $y(0) = 0$ (zero initial condition), the system's total response is its forced response to the shifted step input $\hat{w}(z)$. In particular, the $z$-transform of $w$ can be shown to be

$$\hat{w}(z) = \frac{z^{-1}}{1 - z^{-1}} = \frac{1}{z - 1}$$

and so

$$\hat{y}(z)\left(z + \frac{1}{3}\right) = \frac{1}{z - 1}$$

Solving for $\hat{y}(z)$ gives

$$\hat{y}(z) = \frac{1}{(z - 1)(z + 1/3)}$$

We get the time-domain signal, $y$, from the partial fraction expansion of $\hat{y}(z)$.

$$\hat{y}(z) = \frac{1}{(z - 1)(z + 1/3)} = \frac{K_0}{z + 1/3} + \frac{K_1}{z - 1}$$

Since $\hat{y}(z)$ is a rational function of a complex variable, I can use the residue method to get $K_0$ and $K_1$. In particular,

$$\begin{aligned} K_0 &= \lim_{z \to -1/3} \frac{1}{z - 1} = \frac{1}{-1/3 - 1} = -\frac{3}{4} \\ K_1 &= \lim_{z \to 1} \frac{1}{z + 1/3} = \frac{1}{1 + 1/3} = \frac{3}{4} \end{aligned}$$

So we can see that

$$\begin{aligned} \hat{y}(z) &= -\frac{3}{4}\frac{1}{z + 1/3} + \frac{3}{4}\frac{1}{z - 1} \\ &= -\frac{3}{4}z^{-1}\frac{z}{z + 1/3} + \frac{3}{4}z^{-1}\frac{z}{z - 1} \end{aligned}$$

We use the transforms in table 3 and the time delay operational transform to deduce that

$$\begin{aligned} y(k) &= -\frac{3}{4}\left(-\frac{1}{3}\right)^{k-1} u(k-1) + \frac{3}{4}u(k-1) \\ &= \begin{cases} \frac{3}{4}\left(1 - \left(-\frac{1}{3}\right)^{k-1}\right), & \text{for } k = 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

where $u$ is the discrete-time unit step function. The actual sequence we get is

$$y(0) = 0, y(1) = 0, y(2) = \frac{3}{4} + \frac{1}{4} = 1, y(3) = \frac{3}{4} - \frac{1}{12}, \cdots$$

**Remark:** You can also use PFE by rewriting $\hat{y}(z)$ as a function of $z^{-1}$. This gives

$$\begin{aligned} \hat{y}(z) &= \frac{z^{-1}}{1 - z^{-1}} \frac{z^{-1}}{1 + (1/3)z^{-1}} \\ &= z^{-2} \frac{1}{(1 - z^{-1})(1 + (1/3)z^{-1}} \\ &= z^{-2} \left( \frac{K_1}{1 - z^{-1}} + \frac{K_2}{1 + (1/3)z^{-1}} \right) \end{aligned}$$

Evaluate the residues to get

$$\begin{aligned} K_1 &= \lim_{z^{-1} \to 1} \frac{1}{1 + (1/3)z^{-1}} = \frac{3}{4} \\ K_2 &= \lim_{z^{-1} \to -3} \frac{1}{1 - z^{-1}} = \frac{1}{4} \end{aligned}$$

So we have

$$y(n) = \frac{3}{4} u(n-2) + \frac{1}{4} \left( -\frac{1}{3} \right)^{n-2} u(n-2)$$

The sequence is then

$$y(0) = 0, y(1) = 0, y(2) = 1, y(3) = \frac{3}{4} - \frac{1}{12}, \cdots$$

This is the same we got before, but the expression for the sequence is a bit more compact than what we had before.

**2.3. Frequency Response.** One of the more important features of transfer function representations of a linear system is their relationship to a system's frequency response function. This is easiest to explain for continuous-time LTI systems. In particular, let $\mathbf{G}(s)$ denote a single-input single-output (SISO) transfer function for an LTI system and assume the input is a cosinusoidal input, $w(t) = A\cos(\omega t + \phi)u(t)$ where $A$ is sinusoid's *amplitude*, $\phi$ is the phase angle, and $\omega$ is the frequency. The Laplace transform

of $w$ is

$$W(s) = \frac{A(s\cos\phi - \omega\sin\phi)}{s^2+\omega^2}$$

So the Laplace transform of the output is

$$Y(s) = \mathbf{G}(s)\frac{A(s\cos\phi - \omega\sin\phi)}{s^2+\omega^2}$$

The partial fraction expansion for $Y(s)$ can be written as

$$\begin{aligned} Y(s) &= \frac{K_1}{s-j\omega} + \frac{\overline{K}_1}{s+j\omega} \\ &\quad + \sum \text{terms generated by poles of } \mathbf{G}(s) \end{aligned}$$

If we assume the poles of $\mathbf{G}(s)$ all have negative real parts, then the time-domain terms associated with these poles will asymptotically go to zero as $t \to \infty$. So for large $t$, we can ignore these components and the *steady-state* or *recurrent* behavior of the system becomes

$$Y_{ss}(s) = \frac{K_1}{s-j\omega} + \frac{\overline{K}_1}{s+j\omega}$$

These residues can be shown to have the form

$$K_1 = \frac{1}{2}\mathbf{G}(j\omega)Ae^{j\phi} = \frac{A}{2}|\mathbf{G}(j\omega)|e^{j(\arg G(j\omega)+\phi)}$$

So the steady-state time-domain response of this system is

$$y_{ss}(t) = A|\mathbf{G}(j\omega)|\cos(\omega t + \phi + \arg G(j\omega))$$

This shows that the transfer function allows one to characterize the steady-state response of the LTI system to a sinusoidal input. The amplitude of the response is also a sinusoid of the same frequency with an amplitude equal to $A$ times the modulus of the transfer function $|\mathbf{G}(j\omega)|$. The phase of the output sinusoid equals the sum of the phase angle of the input, $\phi$, and the phase angle of the transfer function, $\arg(\mathbf{G}(j\omega))$. We refer to this pair, $(|\mathbf{G}(j\omega)|, \arg\mathbf{G}(j\omega))$. as the system's *frequency response function*.

Frequency response functions are often used to visualize the behavior of a continuous-time LTI system. This visualization could be obtained

by plotting $|\mathbf{G}(j\omega)|$ and $\arg \mathbf{G}(j\omega)$ as a function of $\omega$. It is more convenient, however, to plot these functions on a log scale. In particular we plot $20 \log_{10} |\mathbf{G}(j\omega)|$ versus $\log \omega$. The particular unit for $20 \log_{10} |\mathbf{G}(j\omega)|$ is called a decibel (dB). We usually plot $\arg \mathbf{G}(j\omega)$ in degrees along the $\log \omega$ axis. Together these plots form the Bode plot of the LTI system. Such plots are relatively easy to plot by hand and are frequently used in designing feedback controllers for a given SISO LTI system.

There is a close relationship between a system's frequency response and its $\mathcal{L}_2$-induced gain. Recall that

$$\|\mathbf{G}\|_{\mathcal{L}_2-\text{ind}} = \inf \left\{ \gamma \in \mathbb{R} \, : \, \|\mathbf{G}[w]\|_{\mathcal{L}_2} \leq \gamma \|w\|_{\mathcal{L}_2} \right\}$$

We can find an explicit formula for this gain when $\mathbf{G}$ is a causal LTI system. In particular, note that Parseval's relation implies

$$\|y\|_{\mathcal{L}_2}^2 \;\; = \;\; \int_{-\infty}^{\infty} |y(\tau)|^2 d\tau = \frac{1}{2\pi} \int_{-\infty}^{\infty} |Y(j\omega)|^2 d\omega$$

where $Y(s)$ is the Laplace transform of $y$. For an LTI system we know $Y(s) = \mathbf{G}(s)W(s)$ where $W(s)$ is the Laplace transform of the input signal. This means we have

$$\begin{aligned} \|y\|_{\mathcal{L}_2}^2 &= \frac{1}{2\pi} \int_{-\infty}^{\infty} |Y(j\omega)|^2 \, d\omega \leq \frac{1}{2\pi} \int_{-\infty}^{\infty} |\mathbf{G}(j\omega)|^2 \, |W(j\omega)|^2 \, d\omega \\ &\leq \left[ \max_{\omega} |\mathbf{G}(j\omega)|^2 \right] \left[ \frac{1}{2\pi} \int_{-\infty}^{\infty} |W(j\omega)|^2 \, d\omega \right] \\ &= \left[ \max_{\omega} |\mathbf{G}(j\omega)|^2 \right] \|w\|_{\mathcal{L}_2}^2 = \|\mathbf{G}\|_{\mathcal{H}_\infty}^2 \, |w|_{\mathcal{L}_2}^2 \end{aligned}$$

This means, therefore that the $\mathcal{L}_2$-induced gain is bounded above by the $\mathcal{H}_\infty$ norm of the system's transfer function $\mathbf{G}$.

$$\|y\|_{\mathcal{L}_2}^2 \leq \|\mathbf{G}\|_{\mathcal{H}_\infty}^2 \, \|w\|_{\mathcal{L}_2}^2$$

As before we can show this upper bound equals the induced gain by finding specific input for which equality holds. Finding such a signal again requires some degree of insight into the system. But we know that $|\mathbf{G}(j\omega)|$ is the magnitude of the output's response when the input is a unit sinusoid

with frequency $\omega$. Since the $\mathcal{H}_\infty$ norm is the largest gain-magnitude of the frequency response function, we simply need to apply a sinusoidal input whose frequency

$$\omega_0 = \arg\max_\omega |\mathbf{G}(j\omega)|$$

The inequality must hold with equality for this particular input and so we can conclude an LTI system's induced gain is simply $\max_\omega |\mathbf{G}(j\omega)| = \|\mathbf{G}\|_{\mathcal{H}_\infty}$.

The preceding formula for the $\mathcal{L}_2$ gain was derived for SISO systems. If $\mathbf{G}$ is a MIMO system, then its induced $\mathcal{L}_2$ gain can be computed in terms of the maximum singular value of the frequency response matrix $\mathbf{G}(j\omega)$. In particular we have

$$\|\mathbf{G}\|_{\mathcal{L}_2-ind} = \sup_{\omega\in\mathbb{R}} \overline{\sigma}(\mathbf{G}(j\omega))$$

These singular values of the matrix $\mathbf{G}(j\omega)$ are denoted as $\sigma(\mathbf{G}(j\omega))$. If we plot these singular values as a function of frequency, we obtain a natural extension of the Bode plot to MIMO systems. In particular, MATLAB's control system toolbox, has a function, `sigma`, that can be used to plot the singular values of a transfer function matrix.

## 3. State Space Realizations

The preceding sections determined the transfer function of a state-space realization $\mathbf{G} \overset{\text{s}}{=} \left[\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right]$ has the form

$$\mathbf{G}(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D}$$

While each state space realization has a unique transfer function, each transfer function has an infinite number of possible state space realizations. This section derives one particular "canonical" realization for a given transfer

function. Note that we can rewrite $\mathbf{G}(s)$ as

$$\mathbf{G}(s) = \frac{Y(s)}{W(s)} = \frac{n(s)}{d(s)} = \frac{b_1 s^{n-1} + b_2 s^{n-2} + \cdots + b_{n-1}s + b_n}{s^n + a_1 s^{n-1} + a_2 s^{n-2} + \cdots + a_{n-1}s + a_n}$$

Let $y^{(k)}$ denote the $k$th time derivative of the signal $y$ with $y^{(0)} = y$. Assume that $y^{(k)}(0) = 0$ for all $0 \leq k \leq n$. A similar set of restrictions will be placed on the input signal $w$. We may then rewrite the above equation as

$$Y(s)(s^n + a_1 s^{n-1} + \cdots + a_{n-1}s + a_n) = W(s)(b_1 s^{n-1} + \cdots + b_{n-1}s + b_n) \tag{16}$$

Take the inverse transform using the fact that the initial values of all derivatives is zero to obtain the $n$th order differential equation

$$y^{(n)} = a_1 y^{(n-1)} + \cdots + a_{n-1}y^{(1)} + a_n y = b_1 u^{(n-1)} + \cdots + b_{n-1}u^{(1)} + b_n u$$

This is an $n$th order differential equation, but we can rewrite it as a set of $n$ first order differential equations.

This is done by introducing an internal *state* variable, $z$, that satisfies the differential equation

$$w(t) = z^{(n)}(t) + a_1 z^{(n-1)}(t) + \cdots + a_{n-1}z^{(1)}(t) + a_n z(t)$$

where $z^{(n)}(t)$ denotes the $n$th time derivative of $z(t)$. Taking the Laplace transform (assuming zero initial conditions) gives

$$W(s) = Z(s)(s^n + a_1 s^{n-1} + \cdots + a_{n-1}s + a_n) \tag{17}$$

We can insert this expression for $W(s)$ into our earlier equation (16) to obtain

$$Y(s) = (b_1 s^{n-1} + \cdots + b_{n-1}s + b_n)Z(s) \tag{18}$$

and taking the inverse transform of equations (17-18) yields

$$w(t) = z^{(n)}(t) + a_1 z^{(n-1)}(t) + \cdots + a_{n-1}z^{(1)}(t) + a_n z(t) \tag{19}$$

$$y(t) = b_1 z^{(n-1)}(t) + \cdots + b_{n-1}z^{(1)}(t) + b_n z(t) \tag{20}$$

Let us now define the following *state variables*, $x_i$ for $i = 1, \ldots, n$, as

$$x_i = z^{(n-i)}$$

for $i = 1, \ldots, n$. This allows us to rewrite equation (19) as

$$\begin{aligned}
\dot{x}_1(t) &= -a_1 x_1(t) - a_2 x_2(t) - \cdots - a_n x_n(t) + w(t) \\
\dot{x}_2(t) &= x_1(t) \\
&\vdots \\
\dot{x}_n(t) &= x_{n-1}(t)
\end{aligned}$$

and we can rewrite equation (20) as

$$y(t) = b_1 x_1(t) + b_2 x_2(t) + \cdots + b_{n-1} x_{n-1}(t) + b_n x_n(t)$$

These state equations in matrix form become

$$\begin{aligned}
\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \\ \dot{x}_3(t) \\ \vdots \\ \dot{x}_{n-1}(t) \\ \dot{x}_n(t) \end{bmatrix} &= \begin{bmatrix} -a_1 & -a_2 & -a_3 & \cdots & -a_{n-1} & -a_n \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \\ \vdots \\ x_{n-1}(t) \\ x_n(t) \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} u(t) \\
&= \mathbf{A}x(t) + \mathbf{B}u(t) \\
y(t) &= \begin{bmatrix} b_1 & b_2 & \cdots & b_{n-1} & b_n \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_{n-1}(t) \\ x_n(t) \end{bmatrix} = \mathbf{C}x(t)
\end{aligned}$$

So what we have done is taken a strictly proper scalar input-output transfer function $\mathbf{G}(s) = \frac{n(s)}{d(s)}$ used it to construct $n$ first order ODEs characterizing the evolution of the state vector, $x$, for the input-output system.

The particular $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ have a convenient form. In particular, the $\mathbf{A}$ matrix is a *companion* matrix. We can readily show that

$$\begin{aligned}\det(\mathbf{A}) &= \det\left(\begin{bmatrix} -a_1 & -a_2 & -a_3 & \cdots & -a_{n-1} & -a_n \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix}\right) \\ &= s^n + a_1 s^{n-1} + a_2 s^{n-2} + \cdots + a_{n-1}s + a_n\end{aligned}$$

In other words, the coefficients of the transfer function matrix' denominator polynomial are embedded in the first row of the $\mathbf{A}$ matrix and that denominator polynomial is equal to $\mathbf{A}$'s characteristic polynomial. The other thing we notice is that the coefficients of transfer function's numerator polynomial are embedded in the $\mathbf{C}$ matrix. These observations suggest that this is a particularly convenient state space realization because it can be written down directly from the transfer function's polynomials. Such "convenient" realizations are said to be *canonical* and what we've shown you above is one form of the companion canonical realization for a transfer function matrix.

Given a transfer function, $\mathbf{G}(s)$, there are actually an infinite number of possible state-space realizations. This can be readily seen as follows. Consider a state space realization (such as the controllable companion form above).

$$\mathbf{G} \stackrel{\text{s}}{=} \left[\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right]$$

Let $\mathbf{Q}$ be any nonsingular square matrix with the same dimensions as $\mathbf{A}$. This means there exists a matrix $\mathbf{Q}^{-1}$ such that $\mathbf{Q}^{-1}\mathbf{Q} = \mathbf{I}$. If $x \in \mathbb{R}^n$ is the state vector for $\mathbf{G}$, we can create a new "state", $z = \mathbf{Q}x$ by passing $x$ through the matrix $\mathbf{Q}$. This means that $x = \mathbf{Q}^{-1}z$ and if we take the time

derivative we get

$$\begin{aligned}\dot{x} &= \mathbf{A}x + \mathbf{B}w \\ &= \mathbf{A}\mathbf{Q}^{-1}z + \mathbf{B}w \\ &= \mathbf{Q}^{-1}\dot{z}\end{aligned}$$

The last two equations can be rewritten as

$$\dot{z} = \mathbf{Q}\mathbf{A}\mathbf{Q}^{-1}z + \mathbf{Q}\mathbf{B}w$$

Note also that

$$\begin{aligned}y &= \mathbf{C}x + \mathbf{D}w \\ &= \mathbf{C}\mathbf{Q}^{-1}z + \mathbf{D}w\end{aligned}$$

This gives rise to the following state space realization

$$\begin{aligned}\dot{z} &= \widetilde{\mathbf{A}}z + \widetilde{\mathbf{B}}w \\ &= \mathbf{Q}\mathbf{A}\mathbf{Q}^{-1}z + \mathbf{Q}\mathbf{B}w \\ y &= \widetilde{\mathbf{C}}z + \widetilde{\mathbf{D}}w \\ &= \mathbf{C}\mathbf{Q}^{-1}z + \mathbf{D}w\end{aligned}$$

where

$$\begin{aligned}\widetilde{\mathbf{A}} &= \mathbf{Q}\mathbf{A}\mathbf{Q}^{-1}, \quad \widetilde{\mathbf{B}} = \mathbf{Q}\mathbf{B} \\ \widetilde{\mathbf{C}} &= \mathbf{C}\mathbf{Q}^{-1}, \quad \widetilde{\mathbf{D}} = \mathbf{D}\end{aligned}$$

We claim that this state space realization, $\left[\begin{array}{c|c} \widetilde{\mathbf{A}} & \widetilde{\mathbf{B}} \\ \hline \widetilde{\mathbf{C}} & \widetilde{\mathbf{D}} \end{array}\right]$ has the same transfer function as the original state space realization, $\left[\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right]$. This assertion is easily verified by directly computing the transfer function of $\left[\begin{array}{c|c} \widetilde{\mathbf{A}} & \widetilde{\mathbf{B}} \\ \hline \widetilde{\mathbf{C}} & \widetilde{\mathbf{D}} \end{array}\right]$

as

$$\begin{aligned}\widetilde{\mathbf{C}}(s\mathbf{I}-\widetilde{\mathbf{A}})^{-1}\widetilde{\mathbf{B}}+\widetilde{\mathbf{D}} &= \mathbf{C}\mathbf{Q}^{-1}(s\mathbf{I}-\mathbf{Q}\mathbf{A}\mathbf{Q}^{-1})^{-1}\mathbf{Q}\mathbf{B}+\mathbf{D}\\ &= \mathbf{C}\mathbf{Q}^{-1}\mathbf{Q}(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{Q}^{-1}\mathbf{Q}\mathbf{B}+\mathbf{D}\\ &= \mathbf{C}(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}+\mathbf{D}\\ &= \mathbf{G}(s)\end{aligned}$$

This verifies that two state space realizations whose state spaces are related through a nonsingular matrix will also have the same transfer function. Since the transfer function characterizes the system's *input/output* behavior, this implies that both state space realizations appear to respond in the same way to the same inputs. So with regard to their input/output behavior the two state space realizations are indistinguishable from each other. This leads to the following theorem

THEOREM 4. *Consider the state-space realization*

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

*and let* $\mathbf{Q}$ *be any nonsingular matrix such that* $z=\mathbf{Q}x$*, then the state space realization*

$$\begin{aligned}\dot{z}(t) &= \mathbf{Q}\mathbf{A}\mathbf{Q}^{-1}z+\mathbf{Q}\mathbf{B}w\\ y(t) &= \mathbf{C}\mathbf{Q}^{-1}z+\mathbf{D}w\end{aligned}$$

*has the transfer function*

$$\mathbf{G}(s)=\mathbf{C}(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}+\mathbf{D}$$

The obvious thing to do is to take a given state space realization $\left[\begin{array}{c|c}\mathbf{A}&\mathbf{B}\\\hline\mathbf{C}&\mathbf{D}\end{array}\right]$ and transform it to a realization $\left[\begin{array}{c|c}\widetilde{\mathbf{A}}&\widetilde{\mathbf{B}}\\\hline\widetilde{\mathbf{C}}&\widetilde{\mathbf{D}}\end{array}\right]$ that is more convenient to work with.

One such "canonical" form is the controller companion form given above. This form is useful because it can be directly written down once we know the transfer function. It is called a controller companion form because it also has an important property known as *controllability* that we will study later. But essentially, this means that we can always find an input $w$ that drives the controller companion form's state to any desired state. Another convenient companion form is the observable companion form

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \vdots \\ \dot{x}_{n-1} \\ \dot{x}_n \end{bmatrix} = \begin{bmatrix} 0 & \cdots & 0 & -\alpha_1 \\ 1 & \cdots & 0 & -\alpha_2 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 1 & -\alpha_{n-1} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_{n-1} \end{bmatrix} w$$

$$y = \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix} + d_0 w$$

which has the transfer function

$$\mathbf{G}(s) = \frac{b_{n-1}s^{n-1} + \ldots + b_1 s + b_0}{s^n + \alpha_{n-1}s^{n-1} + \cdots + \alpha_1 s + \alpha_0} + d_0$$

This form is convenient because it too can be written down as soon as we have an expression for the transfer function and because it has the important property of *observability*, meaning that we can reconstruct the system's initial state from any finite length set of inputs, $w$, and outputs $y$.

Another convenient form is the *modal form*. This one is particularly easy to find when the $\mathbf{A}$ matrix has $n$ distinct eigenvalues $\{\lambda_1, \ldots, \lambda_n\}$ with associated eigenvectors $\{v_1, \ldots, v_n\}$. If we then define the matrices

$$\mathbf{V} = \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix}, \quad \mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{bmatrix}$$

then we can see that

$$\begin{aligned}\mathbf{AV} &= \mathbf{A}\begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix} = \begin{bmatrix} \lambda_1 v_1 & \cdots & \lambda_n v_n \end{bmatrix} \\ &= \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix} \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{bmatrix} \\ &= \mathbf{V}\mathbf{\Lambda}\end{aligned}$$

The matrix $\mathbf{V}$ is invertible and therefore can be used to transform a state space realization $\left[\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right]$ to a modal form.

$$\left[\begin{array}{ccc|c} \lambda_1 & \cdots & 0 & b_0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \lambda_n & b_{n-1} \\ \hline c_0 & \cdots & c_{n-1} & d_0 \end{array}\right]$$

which has the transfer function

$$\mathbf{G}(s) = \sum_{k=0}^{n-1} \frac{c_k b_k}{s - \lambda_{k+1}} + d_0$$

Note that the preceding version of the modal form assumed $\mathbf{A}$ had $n$ distinct eigenvalues. If the matrix is not diagonalizable then some repeated eigenvalues may not have not have enough eigenvectors. In this case, the best we can do is use a similarity transformation to transform $\mathbf{A}$ to its *Jordan Canonical Form* The Jordan Canonical form of $\mathbf{A}$ is

$$\mathbf{J} = \begin{bmatrix} \mathbf{J}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{J}_p \end{bmatrix}$$

where $\mathbf{J}_i$ is an $m_i \times m_i$ matrix of the form

$$\mathbf{J}_i = \lambda_i \mathbf{I}_{m_i} + \mathbf{N}_i$$

in which

$$\mathbf{N}_i = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix}$$

The Jordan canonical form always exists.

## 4. Linearization Methods

In using state-based method we must first have a state-space realization for the system of interest. How are such realizations usually obtained? They can be obtained through a priori mathematical modeling of a physical process or they can be obtained empirically by measuring the system's input/output response. This section will discuss both approaches.

**4.1. A Priori Modeling:** Many mechanical systems have states, $x$, that satisfy the following second order differential equations

$$M(q)\ddot{q} + B(q,\dot{q})\dot{q} + G(q) = F$$

where $q \in \mathbb{R}^n$ is vector of *generalized coordinates*, $M(q)$ is an $n \times n$ non-singular symmetric positive definite matrix called the *Mass matrix*, $F \in \mathbb{R}^n$ is a vector of applied forces (inputs), $G(q)$ is a conservative force vector, and $B(q,\dot{q})$ is a matrix sometimes called the Coriolis/friction matrix. In general, $B(q,\dot{q})$ satisfies

$$\dot{q}^T B(q,\dot{q})\dot{q} \geq 0$$

for all $\dot{q}$.

The dynamics of an inverted pendulum will be used to illustrate how prior modeling can be used to get the preceding differential equation. In

particular, the dynamics of the inverted pendulum

$$m\ell^2\ddot{\theta} = mg\ell\sin\theta - b\dot{\theta} + T$$

where $T$ is a torque applied at the base and $g$ is gravitational acceleration. This equation has the form given above if we let

$$q = \theta, \quad F = T, \quad M(q) = m\ell^2, \quad , B(q) = b, \quad G(q) = -mg\ell\sin\theta$$

A state space realization is a system of first order differential equations, whereas our equations of motion for the inverted pendulum are second order differential equations. We can always rewrite this second order ODE as a set of two first order ODE's through the following state assignments

$$x_1 = q = \theta, \quad x_2 = \dot{q} = \dot{\theta}$$

This then leads to the following state based realization

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= \frac{g}{\ell}\sin x_1 - \frac{b}{m\ell^2}x_2 + \frac{1}{m\ell^2}T \end{aligned}$$

Note that this is NOT a linear differential equation because of the $\sin x_1$ term in the second equation. So this is not a linear state-space realization.

We obtain a "linear" model for the inverted pendulum by linearizing the nonlinear equations. In general, let

$$\dot{x}(t) = f(x(t), w(t))$$

denote the state equation for a given system. We say a state $x^* \in \mathbb{R}^n$ is an *equilibrium* with respect to constant input $\overline{w}$ if $f(x^*, \overline{w}) = 0$. Now assume that the actual input to the system $w(t)$ is a perturbation of the constant input $\overline{w}$. This means there is a function $\delta w$ such that

$$w(t) = \overline{w} + \delta w(t)$$

This perturbation of the input will also generate a perturbation of the state trajectory, $x$ about the equilibrium $x^*$. So there exists a function $\delta x$ such

that

$$x(t) = x^* + \delta x(t)$$

We know that

$$\begin{aligned}
\frac{dx(t)}{dt} &= \frac{d}{dt}x^* + \frac{d}{dt}\delta x(t) = \frac{d}{dt}\delta x(t) \\
&= \left.\frac{\partial f}{\partial x}\right|_{(x^*,\overline{w})} \delta x(t) + \left.\frac{\partial f}{\partial w}\right|_{(x^*,\overline{w})} \delta w(t) + o(|\delta x|) + o(|\delta|) \\
&\approx \left.\frac{\partial f}{\partial x}\right|_{(x^*,\overline{w})} \delta x(t) + \left.\frac{\partial f}{\partial w}\right|_{(x^*,\overline{w})} \delta w(t) \\
&= \mathbf{A}\delta x(t) + \mathbf{B}\delta w(t)
\end{aligned}$$

This forms a linearized state equation about the fixed point $(x^*, \overline{w})$. We call this a Taylor jet linearization of the system about the equilibrium $(x^*, \overline{w})$, since we used a Taylor series expansion for $f$.

Returning to our inverted pendulum example we have an equilibrium when $\theta = x_1 = 0$ and $\dot{\theta} = x_2 = 0$. So $x^* = (0, 0)$. The torque, $T$, needed to ensure this equilibrium is given by

$$0 = \frac{1}{m\ell^2}T$$

or rather $\overline{T} = \overline{w} = 0$. The $f$ function then takes the form

$$f(x, w) = \begin{bmatrix} f_1(x, w) \\ f_2(x, w) \end{bmatrix} = \begin{bmatrix} x_2 \\ \frac{g}{\ell}\sin x_1 - \frac{b}{m\ell^2}x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{1}{m\ell^2} \end{bmatrix} w(t)$$

We can now see that

$$\begin{aligned}
\mathbf{A} &= \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} \end{bmatrix}_{(0,0)} = \begin{bmatrix} 0 & 1 \\ \frac{g}{\ell}\cos x_1^* & -\frac{b}{m\ell^2} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ \frac{g}{\ell} & -\frac{b}{m\ell^2} \end{bmatrix} \\
\mathbf{B} &= \begin{bmatrix} \frac{\partial f_1}{\partial w} \\ \frac{\partial f_2}{\partial w} \end{bmatrix} = \begin{bmatrix} 0 \\ \frac{1}{m\ell^2} \end{bmatrix}
\end{aligned}$$

So our linear state space realization for this system is

$$\mathbf{G} \stackrel{\text{s}}{=} \left[\begin{array}{cc|c} 0 & 1 & 0 \\ \frac{g}{\ell} & -\frac{b}{m\ell^2} & \frac{1}{m\ell^2} \\ \hline 1 & 0 & 0 \end{array}\right]$$

where we took the system's output to be its generalized coordinate $y(t) = q(t) = \theta(t) = x_1(t)$.

Another approach to linearization is by transforming the input, $w$, using a feedback transformation. In particular we will rewrite the input $w$ as

$$w(t) = u(q, \dot{q}) + M(q)v(t)$$

where $u$ is a function of the state, $x = (q, \dot{q})$. We think of this as a reparameterization of the input so that the input becomes $v$, rather than $w$. We will use a particular form for $u$

$$u(q, \dot{q}) = B(q, \dot{q})\dot{q} + G(q)$$

this allows us to rewrite the nonlinear state space equation as

$$M(q)\ddot{q} + B(q, \dot{q})\dot{q} + G(q) = B(q, \dot{q})\dot{q} + G(q) + M(q) = v$$

We can cancel the like terms on both sides of the above equation and take the inverse of the mass matrix $M(q)$ to deduce that

$$\ddot{q} = v$$

This is a second order linear differential equation that can be placed in state space form as

$$\begin{aligned} \dot{x} &= \left[\begin{array}{cc} 0 & I \\ 0 & 0 \end{array}\right] x + \left[\begin{array}{c} 0 \\ I \end{array}\right] v \\ y &= \left[\begin{array}{cc} I & 0 \end{array}\right] x \end{aligned}$$

Essentially, what feedback linearization does is introduce "known state-dependent" terms in the input that cancel out the nonlinear dynamics of the original system, thereby making the entire system appear to be "linear" from the new input $v$ to the output $y$.

Both Taylor and feedback linearizations are used in practice. Taylor linearizations, however, are "approximations" to the original nonlinear system and so they will only be "valid" in a small neighborhood about the chosen equilibrium. For this reason we can think of Taylor linearizations as "local" models of the original system. On the other hand, the feedback linearization is "global". It is not an approximation of the nonlinear system, it holds as long as we know $M(q)$ is invertible, which can be true over a much larger region of the system's state space.

**4.2. Data-Driven Modeling.** Data driven modeling means that we experimentally observe a physical system's input/output behavior and then use those observations to construct a linear state space realization of the process. We will examine two ways of doing this. The first is based on frequency response measurements of the system, which we use to obtain a *transfer function* representation of the system. That transfer function would then be used to construct a state-space realization. The second approach uses time-delayed versions of the output as a *state* for the system and then uses the data matrices generated from inputs and these delay-embedding states to directly identify the state space matrices of the realization.

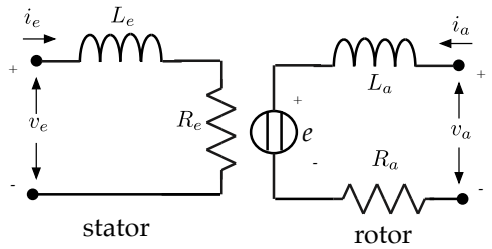



FIGURE 3. DC servo motor

**Frequency-Response Modeling:** In this example we consider a field-controlled DC servo-motor shown in Fig. 3. The state equations associated with the

electrical part of this system are

$$\begin{aligned} v_e &= L_e \frac{di_e}{dt} + R_e i_e \\ v_a &= L_a \frac{di_a}{dt} + R_a i_a + e \end{aligned}$$

where $e$ is the back EMF generated by the motor spinning at angular rate $\omega$. The EMF is proportional to the product of the stator current and the angular rate in which $c$ is the proportionality constant

$$e = c i_e \omega$$

The motor torque is $T = \theta i_e i_a$ where $i_a$ is the rotor current and this defines the mechanical part of the motor. We let $v_e$ be the input to the system, $v$, the output $y = \omega$ is the angular rate, and the other states are $x_1 = i_e$ (stator current), $x_2 = i_a$ (rotor current), and $x_3 = \omega$ (angular rate). With these variable assignments we get the following state equations

$$\begin{aligned} \dot{x}_1 &= -a x_1 + v \\ \dot{x}_2 &= -b x_2 + \rho - c x_1 x_3 \\ \dot{x}_3 &= \theta x_1 x_2 \end{aligned}$$

where $a = R_e/L_e$, $b = R_a/L_a$, and $\rho = v_a/L_a$. The open loop system has an equilibrium at $x_1 = 0$, $x_2 = \rho/b$ and a constant shaft speed setpoint of $\omega_0$.

While we can use the first-principle model to obtain a state equation, this model ignores many of the nonlinearities and other uncertainties in the physical system. In other words, the preceding model is an idealization of an actual DC motor. In this case, it may be better to obtain a model directly from the measured frequency response of the system. Fig. 4 shows the magnitude and phase of the output $y$ for a given unit amplitude sinusoidal input. Each measurement was taken for a unit amplitude input with frequency in column 1. The corresponding max and min output magnitude and phase were then recorded for several separate measurements of that response. The right side of the figure plots these values along with the

max/min error bars. While the variation in the magnitude measurements is relatively small, we see that the phase variation gets very large at the higher frequencies. This phase variation is a consequence of the nonlinearities in the physical system.

| data(:,1) | data(:,2)-data(:,3) | data(:,4)-data(:,5) |
|---|---|---|
| $\omega$(rad/sec) | Magnitude | Phase(deg) |
| $2\pi(.03)$ | 6.6-6.7 | -3 to -5 |
| $2\pi(.1)$ | 6.3-6.5 | -10 to -11 |
| $2\pi(.3)$ | 5.4-5.6 | -23 to -25 |
| $2\pi(1)$ | 2.6-2.65 | -50 to -60 |
| $2\pi(3)$ | 1.12-1.15 | -50 to -70 |
| $2\pi(10)$ | .8-.85 | -55 to -75 |
| $2\pi(30)$ | .32-.5 | -70 to -113 |
| $2\pi(100)$ | .06- .15 | -90 to -130 |
| $2\pi(300)$ | .037-.12 | -89 to -214 |
| $2\pi(1000)$ | .01-.08 | -214 to -326 |

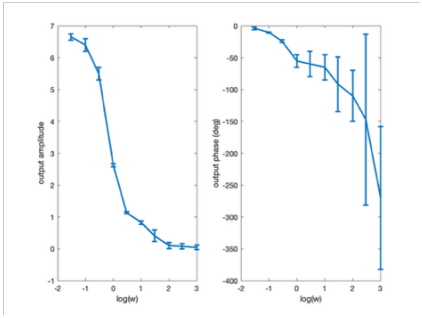



FIGURE 4. Field-Controlled DC Motor Data

We pose an optimization problem that seeks a set of coefficients for a transfer function

$$\mathbf{G}(s) = \frac{b_0 s^2 + b_1 s + b_2}{s^3 + \alpha_0 s^2 + \alpha_1 s + \alpha_2}$$

that minimizes the average $|\log(G(j\omega_i))|$ for each input frequency, $\omega_i$. This can be posed as an optimization function that can be solved using convex programming tools. In matlab we can use the function `fitmagfrd` to do this and thereby obtain the model

$$G(s) = \frac{\frac{20}{3}\left(\frac{s}{5}+1\right)(-s+100)}{\left(\frac{s}{0.5}+1\right)\left(\frac{s}{30}+1\right)(s+100)} = \frac{-20s^2+1900s+10000}{s^3+130.5s^2+3065s+1500}$$

Since we have a strictly proper transfer function, we can readily write out a observable canonical realization for this system.

$$G \overset{s}{=} \left[\begin{array}{ccc|c} 0 & 0 & -1500 & 1 \\ 1 & 0 & -3065 & 0 \\ 0 & 2 & -130.5 & 0 \\ \hline -20 & 1900 & 10000 & 0 \end{array}\right]$$

We can also obtain a modal form for this realization using the eigenvalue/vector decomposition of the $\mathbf{A}$ matrix discussed above

$$G \stackrel{s}{=} \left[\begin{array}{ccc|c} -100 & 0 & 0 & -13.64 \\ 0 & -30 & 0 & -8.275 \\ 0 & 0 & -0.5 & 0.7032 \\ \hline 4 & -3.804 & 4.382 & 0 \end{array}\right]$$

The following Fig. 5 shows the gain/phase magnitude of these continuous-time models against the actual data. These plots show that they are reasonably good fits for both gain and phase.

```
%enter data
data=[0.03 6.6 6.7 -3 -5;
   0.1 6.3 6.5 -10 -11;
   0.3 5.4 5.6 -23 -25;
   1 2.6 2.65 -50 -60;
   3 1.12 1.15 -50 -70;
   10 0.8 0.85 -55 -75;
   30 0.32 0.5 -70 -113;
   100 0.06 0.15 -90 -130;
   300 0.037 0.12 -80 -214;
   1000 0.01 0.08 -214 -326];

w=data(:,1);
mag1=(data(:,2)+data(:,3))/2;
ang1=(data(:,4)+data(:,5))/2;

s = tf('s')
Gx = (-20*s^2+4510*s+10000)/(s^3+130.5*s^2+3065*s+1500);
[g,p,w] = bode(Gx,w);
gr = reshape(g,[10 1]);
pr = reshape(p,[10 1]);
figure(2)
subplot(1,2,1)
plot(log10(w),gr,log10(w),mag1,'o','linewidth',2)
xlabel('log(w)')
ylabel("magnitude")
subplot(1,2,2)
plot(log10(w),pr-360,log10(w),ang1,'o','linewidth',2)
xlabel('log(w)')
ylabel('phase (deg)')
legend("TF","data")
```

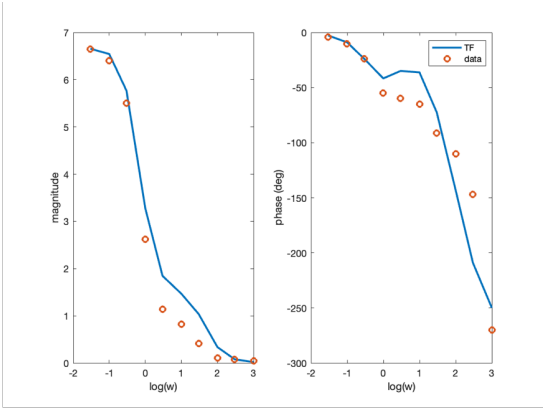



FIGURE 5. Bode Plots of Fit Model

**Dynamic Model Decomposition with Control:** This subsection describes a data-driven way of identifying a state-space realization of a discrete-time dynamical system

$$\begin{aligned} x(k+1) &= \mathbf{A}x(k) + \mathbf{B}w(k) \\ y(k) &= \mathbf{C}x(k) \end{aligned}$$

In the following it will be convenient to represent the state (input or output) at time instant as $x_k$, rather than $x(k)$, with the understanding that $x_k \in \mathbb{R}^n$.

The justification for this approach is based on a well known result from equation-free modeling of continuous-time dynamical systems. This result

essentially says that under certain observability conditions that state information can be determined by simply observing the behavior of the input $w$ and output $y$ over a finite window of time. In other words, we can actually treat the following vector of time-delayed continuous-time output measurements as the state

$$\begin{bmatrix} y(kh) & y((k-1)h) & \cdots & y(0) \end{bmatrix}$$

where $h$ is a regular sampling interval.

In the context of discrete-time systems, we would simply take a sequence of the past outputs as a "surrogate" for the system state. We write the state at time instant $k$ as the vector

$$\mathbf{z}_k = \begin{bmatrix} y_k \\ y_{k-1} \\ \vdots \\ y_{k-N+1} \end{bmatrix}$$

where $y_k$ is the output at time $k$. So our "surrogate" state is a delay-embedded vector of the outputs prior to time $k$ with $N$ denoting the number of outputs we use to form this vector.

To obtain a linear model for the system's dynamics we form the following matrices from the surrogate state vectors $\mathbf{z}_k$ and the known inputs $w_k$

$$\begin{aligned} \mathbf{Z} &= \begin{bmatrix} \mathbf{z}_k & \mathbf{z}_{k-1} & \cdots & \mathbf{z}_{k-M+2} & \mathbf{z}_{k-M+1} \end{bmatrix} \\ \mathbf{Z}^+ &= \begin{bmatrix} \mathbf{z}_{k+1} & \mathbf{z}_k & \cdots & \mathbf{z}_{k-M+1} & \mathbf{z}_{k-M} \end{bmatrix} \\ \mathbf{W} &= \begin{bmatrix} w_k & w_{k-1} & \cdots & w_{k-M+2} & w_{k-M+1} \end{bmatrix} \end{aligned}$$

where $M$ are the number of observations we have made. We can then quickly see that if, on average, the dynamics are linear there should be matrices $\mathbf{A}$ and $\mathbf{B}$ that satisfy

$$\mathbf{Z}^+ = \mathbf{AZ} + \mathbf{BW}$$

In general this will be an overdetermined linear system of linear equations and so we seek a relaxed solution that minimizes

$$\sum_{k=1}^{N-1} \left| \begin{bmatrix} \mathbf{z}_{k+1} \\ w_{k+1} \end{bmatrix} - \begin{bmatrix} \mathbf{A} & \mathbf{B} \end{bmatrix} \begin{bmatrix} \mathbf{z}_k \\ w_k \end{bmatrix} \right|^2$$

which we know can be solved either using singular value decompositions of the data matrices or else through the pseudo-inverse. The resulting "linear" state-space realization is an "average" linear model for the system's dynamics with the average being taking over the window of duration $M$.

We can use this "linear" model as long as $M$ is not too large and as long as we are not in the vicinity of an equilibrium point for the original system. The reason why this is a reasonable approximation is that away from a hyperbolic equilibrium point the "flows" of any smooth dynamical system can be seen as continuous deformations of a "linear" system. This well known result from the elementary theory of ordinary differential equations is known as the Hartman-Grossman Theorem[4].

## 5. Solutions to State Equations

This section discusses the solutions of the state equations. We cover the solutions to continuous-time state equations in some detail. We also cover methods for solving discrete-time state equations.

### 5.1. Solutions to Continuous-time Linear Homogeneous Systems:

Consider the *linear homogeneous* (LH) system whose state trajectory, $x : [t_0, \infty) \to \mathbb{R}^n$, satisfies the initial value problem,

$$\dot{x}(t) = \mathbf{A}(t)x(t), \quad x(t_0) = x_0 \tag{21}$$

[4]The Hartman-Grossman theorem [Hartman (2002)] is a well known result in advanced textbooks on differential equations that essentially says that in the neighborhood of a hyperbolic equilibrium, the flows of the differential equation are topologically equivalent to the flows of its linearization. The theorem is sometimes called the *linearization theorem*

for all $t \geq t_0$. Let $V$ denote the set of all solutions to this problem over a finite interval $[t_0, T]$. Now consider any $\alpha_1, \alpha_2 \in \mathbb{R}$ and $\phi_1, \phi_2 \in V$, then

$$\begin{aligned} \frac{d}{dt}(\alpha_1\phi_1(t) + \alpha_2\phi_2(t)) &= \alpha_1\mathbf{A}(t)\phi_1(t) + \alpha_2\mathbf{A}(t)\phi_2(t) \\ &= \mathbf{A}(t)\left[\alpha_1\phi_1(t) + \alpha_2\phi_2(t)\right] \end{aligned}$$

for all $t$. This means $\alpha_1\phi_1 + \alpha_2\phi_2$ is a solution of the homogeneous problem and is therefore in $V$. So $V$ is closed with respect to addition and dilation and must therefore form a linear space.

We can also show that $V$ has a finite dimension $n$. In particular, let us choose $n$ linearly independent vectors, $\{x_{i0}\}_{i=1}^n$, that span $\mathbb{R}^n$, and let $\{\phi_i\}_{i=1}^n$ denote the solutions to the LH system using initial conditions $\phi_i(t_0) = x_{i0}$. If these solutions were not linearly independent, then there would be scalars $\alpha_i \in \mathbb{R}$ (not all zero) such that for all $t$

$$\sum_{i=1}^{n} \alpha_i\phi_i(t) = 0$$

In particular this would hold at $t_0$, which would contradict the assumption that all $x_{i0}$ are linearly independent. So the solutions $\phi_i$ must be linearly independent and so the dimension of $V$ must be at least $n$. In other words, $\text{span}\{\phi_i\} \subset V$.

To show that $\dim(V) = n$, we need to show that any solution in $V$ lies in the span of the preceding set of $\phi_i$. So let us consider any solution, $\phi$, in $V$ such that $\phi(t_0) = x_0$. We know this $x_0$ lies in the span of $\{x_{i0}\}$, which must also mean that $\phi(t)$ lies in the span of $\{\phi_i\}$. In other words $V \subset \text{span}\{\phi_i\}$ so we can conclude $\text{span}\{\phi_i\} = V$ and since there are $n$ such basis elements, we have $\dim(V) = n$. What we have just shown in summarized in the following theorem

THEOREM 5. *The solutions of the homogeneous system in equation (21) over the interval $[t_0, T]$, form an $n$-dimensional linear space.*

A set , $\{\phi_i\}_{i=1}^n$, of $n$ linearly independent solution of the LH system, $\dot{x} = \mathbf{A}(t)x(t)$ is called a set of *fundamental solutions* and the matrix

$$\mathbf{\Psi}(t) = \left[ \begin{array}{cccc} \phi_1(t) & \phi_2(t) & \cdots & \phi_n(t) \end{array} \right]$$

formed from these fundamental solutions is called a *fundamental matrix* of the LH problem.

Note that any fundamental matrix, $\mathbf{\Psi}$ of the LH problem satisfies the matrix differential equation

$$\dot{\mathbf{\Psi}}(t) = \mathbf{A}(t)\mathbf{\Psi}(t)$$

If $\mathbf{\Psi} = [\phi_1, \ldots, \phi_n]$ is a fundamental matrix and $\phi$ is any solution of the LH problem then there must exist real coefficients (not all zero), $\{\alpha_i\}_{i=1}^n$ such that $\phi(t) = \mathbf{\Psi}(t)\overline{\alpha}$ where $\overline{\alpha} = [\alpha_1, \ldots, \alpha_n]^T$. For any $t$ we know $\phi(t) = \mathbf{\Psi}(t)\overline{\alpha}$ is a linear algebraic equation with a unique solution. We know such LAEs have unique solutions only if the null space of $\mathbf{\Psi}(t)$ is trivial for all $t$, which means that $\mathbf{\Psi}(t)$ must be nonsingular for all $t$. Conversely if $\mathbf{\Psi}$ satisfies the LH matrix differential equation and $\mathbf{\Psi}(t)$ is nonsingular for all $t$, then $\det(\mathbf{\Psi}(t)) \neq 0$ for all $t$. This means the columns of $\mathbf{\Psi}(t)$ are linearly independent for all $t$, which also means $\mathbf{\Psi}(t)$ is a fundamental matrix. What we have just proven is summarized in the following theorem

THEOREM 6. *A solution $\mathbf{\Psi}$ of the matrix differential equation $\dot{\mathbf{\Psi}}(t) = \mathbf{A}(t)\mathbf{\Psi}(t)$ is a fundamental matrix if and only if $\mathbf{\Psi}(t)$ is nonsingular for all $t$.*

Note that any LH problem may have many different fundamental matrices. In practice, we would like a "unique" matrix characterizing the solution of an LH problem. So we pick a specific fundamental matrix whose $i$th column is the fundamental solution generated by the initial state $x(t_0) = e_i$ with $e_i$ being the $i$th elementary basis vector of $\mathbb{R}^n$. We call this particular fundamental matrix a state transition matrix and denote it as $\mathbf{\Phi}(t; t_0)$.

Note that $\mathbf{\Phi}(t;t_0) = \mathbf{\Psi}(t)\mathbf{\Psi}^{-1}(t_0)$ where $\mathbf{\Psi}(t)$ is *any* fundamental matrix of the LH problem. Note that if $\mathbf{T}$ is a nonsingular matrix and we define the matrix $\mathbf{\Psi}_1 = \mathbf{\Psi}_2\mathbf{T}$, then $\mathbf{\Psi}_2$ is a fundamental matrix whenever $\mathbf{\Psi}_1$ is also a fundamental matrix. Moreover, we see that

$$\begin{aligned}\mathbf{\Phi}(t;t_0) &= \mathbf{\Psi}_1(t)\mathbf{\Psi}_1^{-1}(t_0) = \mathbf{\Psi}_2(t)\mathbf{T}\mathbf{T}^{-1}\mathbf{\Psi}_2^{-1}(t_0)\\ &= \mathbf{\Psi}_2(t)\mathbf{\Psi}_2^{-1}(t_0)\end{aligned}$$

This implies the transition matrix is unique and is independent of which fundamental matrix we use to form it.

The following properties of transition matrices are worth itemizing

- $\mathbf{\Phi}(t;t_0)$ is the unique solution to the matrix differential equation

  $$\frac{\partial}{\partial t}\mathbf{\Phi}(t;t_0) = \mathbf{A}(t)\mathbf{\Phi}(t;t_0)$$

  with $\mathbf{\Phi}(t_0;t_0) = \mathbf{I}$.

  **Proof:** For any fundamental matrix $\mathbf{\Psi}$ we know that $\mathbf{\Phi}(t;t_0) = \mathbf{\Psi}(t)\mathbf{\Psi}^{-1}(t_0)$. This implies

  $$\begin{aligned}\frac{\partial}{\partial t}\mathbf{\Phi}(t;t_0) &= \dot{\mathbf{\Psi}}(t)\mathbf{\Psi}^{-1}(t_0)\\ &= \mathbf{A}\mathbf{\Psi}(t)\mathbf{\Psi}^{-1}(t_0) = \mathbf{A}\mathbf{\Phi}(t;t_0). \quad \diamondsuit\end{aligned}$$

- **Group Property:** For all $t, \tau, \sigma \in \mathbb{R}$, then

  $$\mathbf{\Phi}(t;\tau) = \mathbf{\Phi}(t;\sigma)\mathbf{\Phi}(\sigma;\tau)$$

  **Proof:** Note that

  $$\begin{aligned}\mathbf{\Phi}(t;\tau) &= \mathbf{\Psi}(t)\mathbf{\Psi}^{-1}(\tau) = \mathbf{\Psi}(t)\mathbf{\Psi}^{-1}(\sigma)\mathbf{\Psi}(\sigma)\mathbf{\Psi}^{-1}(\tau)\\ &= \mathbf{\Phi}(t;\sigma)\mathbf{\Phi}(\sigma;\tau). \quad \diamondsuit\end{aligned}$$

- $\mathbf{\Phi}(t;t_0)$ is nonsingular for all $t, t_0$ and

  $$[\mathbf{\Phi}(t;t_0)]^{-1} = \mathbf{\Phi}(t_0;t)$$

**Proof:** For any fundamental matrix $\mathbf{\Psi}(t)$, we know that $\det(\mathbf{\Psi}(t)) \neq 0$ for any $t$. This means that

$$\begin{aligned} \det(\mathbf{\Phi}(t;t_0)) &= \det(\mathbf{\Psi}(t)\mathbf{\Psi}^{-1}(t_0)) \\ &= \det(\mathbf{\Psi}(t))\ \det(\mathbf{\Psi}^{-1}(t_0)) \\ &\neq 0, \quad \text{for all } t \end{aligned}$$

So $\mathbf{\Phi}(t;t_0)$ is nonsingular. Finally note that

$$\begin{aligned} [\mathbf{\Phi}(t;t_0)]^{-1} &= \left[\mathbf{\Psi}(t)\mathbf{\Psi}^{-1}(t_0)\right]^{-1} \\ &= \mathbf{\Psi}(t_0)\mathbf{\Psi}^{-1}(t) \\ &= \mathbf{\Phi}(t_0;t). \quad \diamondsuit \end{aligned}$$

- The unique solution $x(t;t_0,x_0)$ to the initial value problem

$$\dot{x}(t) = \mathbf{A}(t)x(t), \quad \text{with } x(t_0) = x_0$$

is

$$x(t;t_0,x_0) = \mathbf{\Phi}(t;t_0)x_0$$

**Proof:** Follows trivially from the above properties. $\diamondsuit$

**5.2. Solutions of Continuous-time Inhomogeneous Problems:** We now extend our earlier solutions to the LH problem to the *inhomogeneous* problem

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

where $u$ is a known input defined over $[t_0,\infty)$. We claim the solution is

$$x(t;t_0,x_0) = \mathbf{\Phi}(t;t_0)x_0 + \int_{t_0}^{t} \mathbf{\Phi}(t;\tau)\mathbf{B}(\tau)u(\tau)d\tau \tag{22}$$

where $\mathbf{\Phi}(t;t_0)$ is the transition matrix for the associated LH problem.

The preceding assertion about the solution to the inhomogeneous problem may be readily verified by direct computation. In particular note that

$$\begin{aligned}
\dot{x}(t;t_0,x_0) &= \dot{\mathbf{\Phi}}(t;t_0)x_0 + \mathbf{\Phi}(t;t)\mathbf{B}(t)u(t) + \int_{t_0}^{t} \dot{\mathbf{\Phi}}(t;\tau)\mathbf{B}(\tau)u(\tau)d\tau \\
&= \mathbf{A}(t)\mathbf{\Phi}(t;t_0)x_0 + \mathbf{B}(t)u(t) + \int_{t_0}^{t} \mathbf{A}(t)\mathbf{\Phi}(t;\tau)\mathbf{B}(\tau)u(\tau)d\tau \\
&= \mathbf{A}(t)\left\{\mathbf{\Phi}(t;t_0)x_0 + \int_{t_0}^{t} \mathbf{\Phi}(t;\tau)\mathbf{B}(\tau)u(\tau)d\tau\right\} + \mathbf{B}(t)u(t) \\
&= \mathbf{A}(t)x(t) + \mathbf{B}(t)u(t)
\end{aligned}$$

Note the important role that the transition matrix plays in characterizing solutions of the inhomogeneous problem. The first term on the right-hand side of equation (22) is associated with the system's zero-input response. The second term is associated with the system's zero-state response. If one wants an explicit solution for $x(t)$, then one needs to find closed-form expressions for the state transition matrix. When the system is LTI so that $\mathbf{A}(t)$ and $\mathbf{B}(t)$ are constant matrices, then there are several approaches one can take to obtain a concrete representation for $\mathbf{\Phi}(t;t_0)$. These approaches are discussed in the next subsection.

**5.3. Solutions to LTI State equations:** Consider the homogeneous LTI system

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

The solution to this ODE can be written as

$$x(t) = x_0 + \int_{t_0}^{t} \mathbf{A}x(\tau)dtau$$

and it can be written as

$$x(t) = \mathbf{\Phi}(t;t_0)x_0$$

It can be shown that this state transition matrix has the form of a matrix exponential function

$$\begin{aligned}\mathbf{\Phi}(t;t_0) &\stackrel{\text{def}}{=} e^{\mathbf{A}(t-t_0)} \\ (23) \qquad &= \mathbf{I} + \mathbf{A}(t-t_0) + \frac{1}{2!}\mathbf{A}^2(t-t_0)^2 + \cdots + \frac{1}{m!}\mathbf{A}^m(t-t_0)^m + \cdots\end{aligned}$$

The derivation of this formula is based on a much deeper result proving that such constant coefficient ODEs always have solutions that are unique. The proof is an interesting application of the contraction mapping principle[5] which shows that any differential equation $\dot{x} = f(x)$ where $f$ satisfies a Lipschitz property has a unique solution [Khalil (2002)]. The proof is constructive and when the system is linear then that construction leads in a natural way to the preceding series formula.

The value of this formula is that there are a number of ways to obtain closed form expressions for the series that we can determine in closed form. We now consider several of these methods.

Consider the inhomogeneous LTI system

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

We know the state trajectory for this system for $t \geq t_0$ is

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

and that the system's output is

$$y(t) = \mathbf{C}e^{\mathbf{A}(t-t_0)} + \int_{t_0}^{t} \mathbf{C}e^{\mathbf{A}(t-\tau)}\mathbf{B}u(\tau)d\tau$$

The last equation's right hand side has two terms. The first term is the natural or zero-input response. The second is the forced or zero-state response

[5] A linear transformation $\mathbf{G} : X \to X$ is a contraction mapping [Antsaklis and Michel (2006)] if $\|G[x-y]\| \leq \|x-y\|$. If $\mathbf{G}$ is a contraction mapping on a complete normed linear space (aka. Banach space) then there exists a unique fixed point $x^*$ in $X$ such that $\mathbf{G}[x^*] = x^*$.

of the system. Since we already know the forced response equals the convolution integral of the input, $u$, with the impulse response function $g$, we can readily see that this system's impulse response function is

$$g(t) = \mathbf{C}e^{\mathbf{A}t}\mathbf{B}u(t)$$

where $u$ is a unit step function in this case and that the Laplace transform of $g$ will be the transfer function which is equal to

$$\mathbf{G}(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}$$

What one might notice in the preceding equations is the important role that the matrix exponential, $e^{\mathbf{A}t}$, plays in all of these equations. So, we need to find methods that can explicitly compute concrete representations for the matrix exponential. This section reviews methods to find $e^{\mathbf{A}t}$.

The matrix exponential in equation (23) is an infinite power series of $\mathbf{A}$. We can use the Cayley-Hamilton theorem to reduce this infinite series to a finite series. Recall that $f(s) = e^{st}$ is an analytic function and that the matrix function $f(\mathbf{A})$ can be written as

$$f(\mathbf{A}) = e^{\mathbf{A}t} = \sum_{k=0}^{\infty} \beta_k \mathbf{A}^k t^k$$

From our earlier work with the Cayley-Hamilton theorem, we know that the division algorithm can be used to write this as

$$e^{\mathbf{A}t} = p(\mathbf{A})q(\mathbf{A}) + r(\mathbf{A})$$

where $p(\mathbf{A}) = \det(s\mathbf{I} - \mathbf{A})$ is the characteristic polynomial of $\mathbf{A}$ and $q(s)$ and $r(s)$ are polynomials with $r(s)$ having a degree less than or equal to $n - 1$. By the Cayley-Hamilton theorem we know $p(\mathbf{A}) = 0$, so that

$$e^{\mathbf{A}t} = r(\mathbf{A}) = \sum_{k=0}^{n-1} \alpha_k(t)\mathbf{A}^k$$

with the coefficients $\alpha_k$ ($k = 0, 1, \ldots, n - 1$) being functions of time that we need to determine.

We can find these functions as follows. For the moment assume that $\mathbf{A}$ has $n$ distinct eigenvalues $i = 1, 2, \ldots, n$, that we denote as $\{\lambda_i\}_{i=1}^n$. These eigenvalues satisfy the characteristic equation so that $p(\lambda_i) = 0$ for $i = 1, 2, \ldots, n$. We also know by the division algorithm that

$$f(\lambda_i) = e^{\lambda_i t} = p(\lambda_i)q(\lambda_i) + r(\lambda_i)$$

where $r(\lambda_i)$ is a polynomial of degree $n - 1$ or less. Using the fact that $p(\lambda_i) = 0$, this reduces to

$$\begin{aligned} e^{\lambda_i t} &= r(\lambda_i) \\ &= \sum_{k=0}^{n-1} \alpha_k(t)\lambda_i^k \end{aligned}$$

since we have $n$ distinct $\lambda_i$, this gives a set of $n$ linear equations that we can then solve for the coefficient functions, $\alpha_k(t)$ ($k = 0, 1, \ldots, n-1$).

**Example:** Consider the matrix $\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix}$. The characteristic polynomial of $\mathbf{A}$ is

$$p(s) = s^2 + 3s + 2 = 0$$

So the eigenvalues of $\mathbf{A}$ are $\lambda_1 = -1$ and $\lambda_2 = -2$. The reduced order representation for $e^{\mathbf{A}t}$ is

$$e^{\mathbf{A}t} = \alpha_0(t)\mathbf{I} + \alpha_1(t)\mathbf{A}$$

To get these coefficient functions, we use the eigenvalues of $\mathbf{A}$ to form the algebraic equations

$$\begin{aligned} e^{-t} &= \alpha_0(t) - \alpha_1(t) \\ e^{-2t} &= \alpha_0(t) - 2\alpha_1(t) \end{aligned}$$

which we can solve to obtain

$$\begin{aligned} \alpha_0(t) &= 2e^{-t} - e^{-2t} \\ \alpha_1(t) &= e^{-t} - e^{-2t} \end{aligned}$$

and so our reduced order expression for the matrix exponential is

$$\begin{aligned} e^{\mathbf{A}t} &= (2e^{-t} - e^{-2t})\mathbf{I} + (e^{-t} - e^{-2t})\mathbf{A} \\ &= \begin{bmatrix} 2e^{-t} - e^{-2t} & e^{-t} - e^{-2t} \\ -2e^{-t} + 2e^{-2t} & -e^{-t} + 2e^{-2t} \end{bmatrix} \end{aligned}$$

**Example:** Here is an example where the eigenvalues of $\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$ are not real. The characteristic equation is $s^2 + 1$ which has eigenvalues $\lambda_1 = j$ and $\lambda_2 = -j$. So the equations we need to solve for the coefficients in

$$e^{\mathbf{A}t} = \alpha_0(t)\mathbf{I} + \alpha_1(t)\mathbf{A}$$

are now

$$\begin{aligned} e^{jt} &= \cos(t) + j\sin(t) = \alpha_0(t) + \alpha_1(t)j \\ e^{-jt} &= \cos(t) - j\sin(t) = \alpha_0(t) - \alpha_1(t)j \end{aligned}$$

Solving the $\alpha_0$ and $\alpha_1$ gives

$$\begin{aligned} \alpha_0(t) &= \cos(t) \\ \alpha_1(t) &= \sin(t) \end{aligned}$$

so that

$$\begin{aligned} e^{\mathbf{A}t} &= \cos(t)\mathbf{I} + \sin(t)\mathbf{A} \\ &= \begin{bmatrix} \cos(t) & \sin(t) \\ -\sin(t) & \cos(t) \end{bmatrix} \end{aligned}$$

Note that if one or more of the eigenvalues is repeated, the above procedure will not generate $n$ linearly independent equations. For any eigenvalue with multiplicity $m > 1$, however, we know the first $(m-1)$ derivatives of $p(s)$ must vanish at the eigenvalues. We can use this fact to generate

additional linearly independent equations. So if $\lambda_i$ is an eigenvalue with multiplicity $m_i > 1$, then

$$\begin{aligned} f(\lambda_i) &= \sum_{k=0}^{n-1} \alpha_k \lambda_i^k = r(\lambda_i) \\ f^{(1)}(\lambda_i) &= r^{(1)}(\lambda_i) \\ &\vdots \\ f^{(m_i-1)}(\lambda_i) &= r^{(m_i-1)}(\lambda_i) \end{aligned}$$

where $f^{(j)}(s)$ and $r^{(j)}(s)$ denote the $j$th derivatives of $f(s)$ and $r(s)$, respectively. These additional equations will generate $m_i$ linearly independent equations and since $\sum_i m_i = n$, we will be able to generate enough equations to characterize all $\alpha_k(t)$ for $k = 0, 1, 2, \ldots, n-1$.

**Example:** Find the matrix exponential for $\mathbf{A} = \begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix}$. This matrix has an eigenvalue $\lambda_1 = 0$ with multiplicity $m_1 = 2$. We need to find $\alpha_0(t)$ and $\alpha_1(t)$ such that

$$f(\mathbf{A}) = e^{\mathbf{A}t} = \alpha_0(t)\mathbf{I} + \alpha_1(t)\mathbf{A} \stackrel{\text{def}}{=} r(\mathbf{A})$$

The equations we need to solve for these coefficients are

$$\begin{aligned} e^{\lambda_i t} = f(\lambda_1) &= \alpha_0(t) + \alpha_1(t)\lambda_1 = r(\lambda_1) \\ f^{(1)}(\lambda_1) = \left.\frac{d}{ds}e^{st}\right|_{s=\lambda_1} &= te^{\lambda_1 t} \\ &= r^{(1)}(\lambda_1) = \left.\frac{dr(s)}{ds}\right|_{s=\lambda_1} = \alpha_1(t) \end{aligned}$$

Since $\lambda_1 = 0$, these two equations become

$$\begin{aligned} 1 &= \alpha_0(t) \\ t &= \alpha_1(t) \end{aligned}$$

and so

$$\begin{aligned} e^{\mathbf{A}t} &= \mathbf{I} + t\mathbf{A} \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} -t & t \\ -t & t \end{bmatrix} \\ &= \begin{bmatrix} 1-t & t \\ -t & 1+t \end{bmatrix} \end{aligned}$$

**Other approaches - truncation:** The matrix exponential may also be approximated by truncating the power series. In general, the power series converges slowly and so one would need many many terms for a useable approximation. In some cases, the $\mathbf{A}^k$ term map be zero far out in the sequence [6]. For example in our earlier example where $\mathbf{A} = \begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix}$, we see that $\mathbf{A}$ is nilpotent since $\mathbf{A}^2 = 0$. The matrix exponential power series can therefore be written as

$$\begin{aligned} e^{\mathbf{A}t} &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix} t \\ &= \mathbf{I} + \mathbf{A}t = \begin{bmatrix} 1-t & t \\ -t & 1+t \end{bmatrix} \end{aligned}$$

which is the same answer we got above.

**Other approaches - Laplace Transforms:** Another way to find the matrix exponential is to use Laplace transforms. Note that $e^{\mathbf{A}t}$ satisfies the matrix differential equation

$$\frac{d}{dt} e^{\mathbf{A}t} = \mathbf{A} e^{\mathbf{A}t}, \quad e^{\mathbf{A}0} = \mathbf{I}$$

Since $\mathbf{e}^{\mathbf{A}t}$ is defined for all values of $t \in \mathbb{R}$, we take the bilateral Laplace transform of $e^{\mathbf{A}t}$. Let $\hat{\mathbf{\Phi}}(s)$ denote the bilateral Laplace transform of $\mathbf{\Phi}(t) =$

[6] A matrix $\mathbf{A}$ for which $\mathbf{A}^k = 0$ when $k$ is a positive integer is said to be *nilpotent*. The matrix is said to be *idempotent* if $\mathbf{A}^2 = \mathbf{A}$.

$e^{\mathbf{A}t}$. Taking the Laplace transform of the preceding matrix differential equation yields,

$$s\hat{\mathbf{\Phi}}(s) - e^{\mathbf{A}t}\big|_{t=0} = \mathbf{A}\hat{\mathbf{\Phi}}(s)$$

Solving for $\hat{\mathbf{\Phi}}(s)$ gives

$$\hat{\mathbf{\Phi}}(s) = (s\mathbf{I} - \mathbf{A})^{-1}$$

So we can get a concrete representation for $e^{\mathbf{A}t}$ by taking the inverse transform of the Laplace transform of the expression (usually done using a partial fraction expansion).

**Example:** Let $\mathbf{A} = \begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix}$, then

$$s\mathbf{I} - \mathbf{A} = \begin{bmatrix} s+1 & -1 \\ 1 & s-1 \end{bmatrix}$$

and

$$(s\mathbf{I} - \mathbf{A})^{-1} = \frac{1}{s^2}\begin{bmatrix} s-1 & 1 \\ -1 & s+1 \end{bmatrix}$$

Taking the inverse transform gives

$$\mathcal{L}^{-1}\begin{bmatrix} \frac{1}{s} - \frac{1}{s^2} & \frac{1}{s^2} \\ -\frac{1}{s^2} & \frac{1}{s} + \frac{1}{s^2} \end{bmatrix} = \begin{bmatrix} 1-t & t \\ -t & 1+t \end{bmatrix}$$

In general if we use partial fraction expansions of $(s\mathbf{I} - \mathbf{A})^{-1}$ to find the matrix exponential, we get an expression of the following form,

$$e^{\mathbf{A}t} = \sum_{i=1}^{\sigma} \sum_{k=0}^{m_i - 1} \mathbf{A}_{ik} t^k e^{\lambda_i t}$$

where the residues are

$$\mathbf{A}_{ik} = \frac{1}{k!} \frac{1}{(m_i - 1 - k)!} \lim_{s \to \lambda_i} \left\{ \frac{d^{m_i - 1 - k}}{ds^{m_i - 1 - k}} \left[ (s - \lambda_i)^{m_i} (s\mathbf{I} - \mathbf{A})^{-1} \right] \right\}$$

with $\mathbf{A}$ having $\sigma$ distinct eigenvalues and the $i$th distinct eigenvalue being denoted as $\lambda_i$ with multiplicity $m_i$. The residues are called the *modes* of

the system. Sometimes, one refers to the entire term, $\mathbf{A}_{ik}t^k e^{\lambda_i t}$, as a system *mode*.

From the preceding equation, one can readily see that if $\mathrm{Re}(\lambda_i) < 0$ for all $i$ then $e^{\mathbf{A}t} \to 0$ as $t \to \infty$. We will later say that this means the origin of the LTI system is *asymptotically stable*. Note that this eigenvalue condition is a necessary and sufficient condition for the asymptotic stability of the LTI system's origin.

**Other approaches - diagonalization:** We may also use similarity transformations to compute the matrix exponential. This is particularly useful when $\mathbf{A}$ is diagonalizable through a nonsingular matrix $\mathbf{T}$. In this case,

$$\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n) = \mathbf{T}^{-1}\mathbf{A}\mathbf{T}$$

and we see that

$$e^{\mathbf{\Lambda}t} = e^{\mathbf{T}^{-1}\mathbf{A}\mathbf{T}t} = \mathrm{diag}(e^{\lambda_1 t}, \ldots, e^{\lambda_n t})$$

and so the transition matrix becomes

$$\mathbf{\Phi}(t) = e^{\mathbf{A}t} = \mathbf{T}\mathrm{diag}(e^{\lambda_i t})\mathbf{T}^{-1}$$

If $\mathbf{A}$ is not diagonalizable, then one uses its Jordan canonical form to compute the matrix exponential. In this case, we reduce the system to

$$\mathrm{diag}(\mathbf{J}_i) = \mathbf{T}^{-1}\mathbf{A}\mathbf{T}$$

where the $i$th Jordan block is an $m_i \times m_i$ square matrix

$$\mathbf{J}_i = \begin{bmatrix} \lambda_1 & 1 & 0 & \cdots & 0 \\ 0 & \lambda_i & 1 & \cdots & 0 \\ 0 & 0 & \lambda_i & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_i \end{bmatrix}$$

where $\lambda_i$ is the $i$th distinct eigenvalue with multiplicity $m_i$. For this case, one can verify by direct computation that

$$e^{\mathbf{J}_i t} = e^{\lambda_i t} \begin{bmatrix} 1 & t & \cdots & \frac{t^{m_i-1}}{(m_i-1)!} \\ 0 & 1 & \cdots & \frac{t^{m_i-2}}{(m_i-2)!} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

and the matrix exponential is

$$e^{\mathbf{A}t} = \mathbf{T}\mathrm{diag}\left(e^{\mathbf{J}_i t}\right)\mathbf{T}^{-1}$$

**5.4. Discrete-time Transition Matrix.** Consider the discrete-time state space equations

$$\begin{aligned} x(k+1) &= \mathbf{A}(k)x(k) + \mathbf{B}(k)u(k) \\ y(k) &= \mathbf{C}(k)x(k) + \mathbf{D}(k)u(k) \end{aligned}$$

We want to find solutions $y(k)$ and $x(k)$ for all $k \geq k_0$ for a given $u$ assuming $x(k_0) = x_0$.

As before we start with the homogenous problem,

$$x(k+1) = \mathbf{A}(k)x(k)$$

and observe that

$$\begin{aligned} x(k+2) &= \mathbf{A}(k+1)x(k+1) \\ &= \mathbf{A}(k+1)\mathbf{A}(k)x(k) \\ &\vdots \\ x(k+n) &= \mathbf{A}(n-1)\mathbf{A}(n-2)\cdots\mathbf{A}(k+1)\mathbf{A}(k)x(k) \\ &= \prod_{j=k}^{n-1}\mathbf{A}(j)x(k) \end{aligned}$$

This suggests that the transition matrix is

$$\mathbf{\Phi}(n;k) = \prod_{j=k}^{n-1}\mathbf{A}(j), \quad n > k$$

and that $\mathbf{\Phi}(k;k) = \mathbf{I}$. So the solution to the homogeneous problem is

$$x(n) = \mathbf{\Phi}(n;k_0)x_{k_0} = \prod_{j=k_0}^{n-1} \mathbf{A}(j)x(k_0)$$

for all $n > k_0$.

Computing closed form expressions for $\mathbf{\Phi}(n;k_0)$ may be difficult in practice. One way of doing this is to form the first few matrices $\mathbf{\Phi}(0,0)$, $\mathbf{\Phi}(1;0)$, $\mathbf{\Phi}(2,0)$, and so on to see if a "pattern" emerges. One would then use that proposed "pattern" as a propositional statement on the set of natural numbers whose satisfaction would need to be formally verified using the principle of mathematical induction.

Some common properties for continuous time transition matrices are similar to those for discrete-time systems. For instance the discrete time transition matrix has the semigroup property

$$\mathbf{\Phi}(k;\ell) = \mathbf{\Phi}(k;m)\mathbf{\Phi}(m;\ell), \quad k \geq m \geq \ell$$

But not all properties of continuous-time transition matrices carry over to discrete time. In particular, recall that for continuous time, if $t > \tau$, then future value of the state at $t$ are obtained from past values *and* vice versa. In other words, the continuous-time state transition matrix is invertible so that time moves freely in both directions. This is not the case for discrete-time systems unless $\mathbf{A}^{-1}(k)$ exists for all $k$.

If we use partial fractions of $(z\mathbf{I} - \mathbf{A})^{-1}$ to find the discrete-time transition matrix we get

$$\begin{aligned}\mathbf{A}^k &= \sum_{i=1}^{\sigma} \big(\mathbf{A}_{i,0}\lambda_i^k u_k + \mathbf{A}_{i,1}k\lambda_i^{k-1}u_{k-1} + \\ &\quad \cdots + \mathbf{A}_{i,n_i-1}k(k-1)\cdots(k-n_i+2)\lambda_i^{k-n_i+1}u_{k-n_i+1}\big)\end{aligned}$$

where $\sigma$ is the number of distinct eigenvalues of $\mathbf{A}$, $n_i$ is the multiplicity of the $i$th distinct eigenvalue and the residues are

$$\mathbf{A}_{i,\ell} = \frac{1}{\ell!}\frac{1}{(n_i-1-\ell)!}\lim_{z\to\lambda_i}\left\{\frac{d^{n_i-\ell-1}}{dz^{n_i-\ell-1}}(z-\lambda_i)^{n_i}(z\mathbf{I}-\mathbf{A})^{-1}\right\}$$

Note that if $|\lambda_k| < 1$ for all $i$, then $\mathbf{A}^k \to 0$ as $k \to \infty$, which corresponds to asymptotic stability of the origin of the discrete time LTI system. This eigenvalue condition is again necessary and sufficient the origin's asymptotic stability in discrete-time LTI systems.

**Example:** Consider the system

$$x(k+1) = \begin{bmatrix} 1 & (k+1) \\ 0 & 1 \end{bmatrix} x(k)$$

for $k \geq 0$. Determine the system's state transition matrix.

So we first compute the first few state transition matrices,

$$\begin{aligned}
\mathbf{\Phi}(0,0) &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \\
\mathbf{\Phi}(1,0) &= \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \\
\mathbf{\Phi}(2,0) &= \mathbf{\Phi}(2,1)\mathbf{\Phi}(1,0) \\
&= \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \\
&= \begin{bmatrix} 1 & 1+2 \\ 0 & 1 \end{bmatrix} \\
\mathbf{\Phi}(3,0) &= \mathbf{\Phi}(3,2)\mathbf{\Phi}(2,0) \\
&= \begin{bmatrix} 1 & 3 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 1+2 \\ 0 & 1 \end{bmatrix} \\
&= \begin{bmatrix} 1 & 1+2+3 \\ 0 & 1 \end{bmatrix}
\end{aligned}$$

The preceding computations suggest a "pattern" which is

$$\mathbf{\Phi}(k,0) = \begin{bmatrix} 1 & \sum_{j=1}^{k} j \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & \frac{k(k+1)}{2} \\ 0 & 1 \end{bmatrix}$$

for $k \geq 1$. The truth of this proposition is easily verified using mathematical induction. In particular, the base step would require

$$\mathbf{\Phi}(1,0) = \begin{bmatrix} 1 & \frac{1\times 2}{2} \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

which is clearly true. For the inductive step, we assume $\mathbf{\Phi}(k;0) = \begin{bmatrix} 1 & \frac{k(k+1)}{2} \\ 0 & 1 \end{bmatrix}$ and then consider

$$\begin{aligned} \mathbf{\Phi}(k+1;0) &= \mathbf{\Phi}(k+1;k)\mathbf{\Phi}(k;0) \\ &= \begin{bmatrix} 1 & k+1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & \frac{k(k+1)}{2} \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & \frac{k(k+1)}{2} + k + 1 \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & \frac{k(k+1)+2(k+1)}{2} \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & \frac{(k+1)(k+2)}{2} \\ 0 & 1 \end{bmatrix} \end{aligned}$$

which clearly verifies the inductive step and so the proposition must hold for all $n \in \mathbb{N}$. The actual transition matrix is

$$\begin{aligned} \mathbf{\Phi}(k;\ell) &= \mathbf{\Phi}(k,0)\mathbf{\Phi}^{-1}(\ell,0) \\ &= \begin{bmatrix} 1 & \frac{k(k+1)}{2} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & -\frac{\ell(\ell+1)}{2} \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & \frac{k(k+1)-\ell(\ell+1)}{2} \\ 0 & 1 \end{bmatrix} \end{aligned}$$

CHAPTER 3

# Stability

**Stability** is one of the most important concepts used in describing a dynamical system's qualitative behavior. This chapter examines Lyapunov stability for linear systems and develops the $\mathcal{L}_p$ notion of input/output stability. Much of this material is drawn from [Khalil (2002)].

## 1. Lyapunov Stability

Consider a time-invariant system whose state trajectory, $x : \mathbb{R} \to \mathbb{R}^n$ satisfies the initial value problem (IVP)

$$\dot{x}(t) = f(x(t)), \quad x(0) = x_0 \tag{24}$$

where $f : D \to \mathbb{R}^n$ is locally Lipschitz[1] in an open connected set $D \subset \mathbb{R}^n$ (also called the "domain"). The Lipschitz condition is required to ensure the existence of unique local solutions about the initial state, $x_0$.

Given the IVP in equation (24), we say a point $x^* \in D$ is an *equilibrium point* of the system if $f(x^*) = 0$. We may assume without loss of generality that $x^* = 0$, since if $x^* \neq 0$ we can introduce a change of variables $z = x - x^*$ whose differential equation

$$\dot{z} = \frac{d}{dt}(x - x^*) = \dot{x} = f(x(t)) = f(z(t) + x^*)$$

has an equilibrium point at the origin. For this reason, we assume without loss of generality (wlog) that the equilibrium is always at the origin.

[1] A function $f : \mathbb{R}^n \to \mathbb{R}^n$ is locally Lipschitz at $x \in \mathbb{R}^n$ if there is a non-negative constant $L > 0$ and a neighborhood about $x$ such that for any $z, y$ in this neighborhhood we have $|f(y) - f(z)| \leq L|y - z|$.

We say that the equilibrium, $x^* = 0$, is *stable in the sense of Lyapunov* if for all $\epsilon > 0$ there exists $\delta > 0$ such that if $|x(0)| < \delta$, then $|x(t)| < \epsilon$ for all $t \geq 0$. We say the equilibrium is *unstable* if it is not stable. We say the equilibrium is *asymptotically stable* if it is stable and $x(t) \to 0$ as $t \to \infty$ for all $x(0)$ in a neighborhood[2] of the origin.

Note that Lyapunov stability is a *local* property of the equilibrium since it is defined in a neighborhood, $N_\delta(x^*)$, of the equilibrium. In particular, we see that the initial state, $x(0)$, must start within a distance $\delta$ to ensure that it remains in a distance $\epsilon$ of the origin for all time. The choice of $\epsilon$ is arbitrary, and this means $\delta$ is a function of $\epsilon$; that we denote as $\delta(\epsilon)$. It is possible for $\lim_{\epsilon \to \infty} \delta(\epsilon) =$ constant which would mean that Lyapunov stability can only be assured if $x(0)$ starts within this constant distance, $\delta$, from the origin (hence the local nature of the concept). If one can show that $\delta(\epsilon) \to \infty$ as $\epsilon \to \infty$, then the stability concept would be a *global* property of the equilibrium.

The notion of Lyapunov stability may be interpreted as generalizing the intuitive notion that "stable" systems dissipate energy. Energy is a non-negative scalar function of state and a system that dissipates energy would have its energy decreasing over time. This suggests that one can certify whether the equilibrium is stable or asymptotically stable by finding a scalar function, $V : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$, acting as a "surrogate" for the system's energy that is always decreasing along the system's trajectories. This notion of the surrogate energy function is formalized in the following theorem.

THEOREM 7. **Lyapunov's Direct Method:** *Let* $0$ *be the equilibrium point for the system* $\dot{x}(t) = f(x(t))$ *where* $f : D \to \mathbb{R}^n$ *is locally Lipschitz on the connected open set,* $D \subset \mathbb{R}^n$. *If there exists a* $C^1$ *function* $V : D \to \mathbb{R}$ *such that*

[2]Let $p \in \mathbb{R}^n$, then for some $\epsilon > 0$, an $\epsilon$-neighborhood of $p$ is $N_\epsilon(p) = \{q \in \mathbb{R}^n : |p - q| < \epsilon\}$. Note that $N_\epsilon(p)$ is an open set in the sense that for any $\epsilon > 0$ there is another point $q \in N_\epsilon(p)$. . The closure of $N_\epsilon(p)$ contains all limit points of $N_\epsilon(p)$ and is denoted as $\overline{N_\epsilon(P)}$ [Rudin (1964)].

- $V$ *is* positive definite*; i.e.* $V(0) = 0$ *and* $V(x) > 0$ *for all* $x \in D - \{0\}$,
- $\dot{V}(x) = \dfrac{\partial V(x)}{\partial x} f(x)$ *is* negative semidefinite*; i.e.* $\dot{V}(x) \leq 0$ *for all* $x \in D$,

*then* $x = 0$ *is* stable *in the sense of Lyapunov. Furthermore if one can show that* $\dot{V}(x)$ *is* negative definite *(i.e.* $\dot{V}(0) = 0$ *and* $\dot{V}(x) < 0$ *for all* $x \in D - \{0\}$*), then the equilibrium is* asymptotically stable.

**Proof:** For any $\epsilon > 0$, we need to find a starting neighborhood for which the system state remains within a distance $\epsilon$ of the origin for all future time. So consider the open ball,

$$N_\epsilon(0) = \{x \in \mathbb{R}^n \, : \, |x| < \epsilon\}$$

and consider $V(x)$ evaluated on the boundary of this neighborhood

$$\partial N_\epsilon(0) = \{x \in \mathbb{R}^n \, : \, |x| = \epsilon\}$$

This set is *closed* and *bounded*[3] and so is *compact*. A well known fact from real analysis [Rudin (1964)] is that all continuous functions attain their maximum and minimum on compact sets. So there exists a real number $\alpha \geq 0$ such that

$$\alpha = \min_{x \in \partial N_\epsilon(0)} V(x)$$

Now let $\Omega_\alpha$ be a subset of $\mathbb{R}^n$ such that

$$\Omega_\alpha = \{x \in N_\epsilon(0) \, : \, V(x) < \alpha\}$$

This set is contained in $N_\epsilon(0)$.

[3] A set $M$ in a normed linear space is closed if it contains its limit points and it is bounded if there exists $R > 0$ such that $\|x - y\| \leq R$ for all $x, y \in M$. When $M$ is a closed and bounded subset of $\mathbb{R}^n$, then it is said to be *compact*.

So let $x$ be any solution to $\dot{x}(t) = f(x(t))$ where $x(0) \in \Omega_\alpha$. By assumption we know for all $x \in D$ that

$$\dot{V}(x) = \frac{\partial V(x)}{\partial x} f(x) \leq 0$$

which means $V(x(t))$ is a monotone non-increasing function of time and so

$$V(x(t)) \leq V(x(0)) < \alpha$$

for all $t \geq 0$. In other words $x(t)$ remains in $\Omega_\alpha$ for all $t \geq 0$ (such a set is said to be *forward invariant*). Since $V$ is continuous and $V(0) = 0$, there must exist a $\delta > 0$ such that $|x| < \delta$ implies $V(x) < \alpha$, so we can define another open ball

$$N_\delta(0) = \{x \in \mathbb{R}^n \,:\, |x| < \delta\}$$

We can therefore see that $N_\delta(0) \subset \Omega_\alpha \subset N_\epsilon(0)$. So if we take $x(0) \in N_\delta(0)$, then we know $x(t) \in \Omega_\alpha$ for all $t$ (by forward invariance) and so $x(t) \in N_\epsilon(0)$ for all $t$, which is precisely the definition of Lyapunov stability.

To prove the assertion regarding asymptotic stability, we note that if $\dot{V}(x)$ is negative definite, then $V(x(t))$ is a monotone decreasing function. Since $V(x) \geq 0$ we know this decreasing function is bounded below by 0 and so by the bounded monotone convergence theorem in real analysis[4]. We know there exists $c \geq 0$ such that $V(x(t)) \to c$ as $t \to \infty$. If $c$ is strictly greater than zero, then the continuity of $V$ would mean there is a $d > 0$ such that

$$N_d(0) = \{x \in \mathbb{R}^n \,:\, |x| < d\} \subset \Omega_c = \{x \in \mathbb{R}^n \,:\, V(x) < c\}$$

The trajectory, $x(t)$, cannot enter $N_d(0)$, since $V(x(t))$ is always greater than $c$. So let

$$-\gamma = \max_{x \in \overline{\Omega}_c - N_d(0)} \dot{V}(x) < 0$$

[4]Let $\{x_i\}_{i=1}^\infty$ denote a sequence of real numbers. If there exists $B \in \mathbb{R}$ such that $x_i \geq B$ (sequence is *bounded*) for all $i$ and $x_{i+1} \leq x_i$ for all $i$ (sequence is *monotone decreasing*), then the sequence is convergent to a limit $x^* \geq B$.

which must be strictly greater than zero since $c > 0$ and $\dot{V}(x)$ is negative definite (i.e. only zero at zero). This implies

$$V(x(t)) = V(x(0)) - \int_0^t \dot{V}(x(\tau))d\tau \leq V(x(0)) - \gamma t$$

This last equation, however, implies that if $t$ is large enough, i.e. when $t > \frac{V(x(0))}{\gamma}$, then $V(x(t))$ will be negative which cannot occur since we know $V(x)$ is positive definite. So $c$ cannot be positive, it must be zero and so the state trajectory $x(t)$ asymptotically converges to zero. $\diamondsuit$

A $C^1$ function $V : D \to \mathbb{R}$ that satisfies the conditions in Lyapunov's Direct Method is called a *Lyapunov function*. In recent years, it has also been referred to as a *certificate* of Lyapunov stability since the "existence" of the function is sufficient to "certify" that the origin is Lyapunov stable.

In general, finding Lyapunov functions can be difficult to do. We usually take a function that is known to be a Lyapunov certificate for a system related to the one whose stability we want to verify and then parameterize that known certificate in a manner that allows us to search for a Lyapunov certificate of the system we're interested in. This is commonly done in certifying the stability of the origin of a nonlinear system

$$\dot{x}(t) = f(x(t))$$

In this case, we would take the Lyapunov function for its Taylor linearization about the equilibrium

$$\dot{x}(t) = \left.\frac{\partial f}{\partial x}\right|_{x=0} x(t) = \mathbf{A}x(t)$$

As we will show below, the Lyapunov function for the linearization is a function $V : D \to \mathbb{R}^n$ that takes values

$$V(x) = x^T \mathbf{P} x \tag{25}$$

where $\mathbf{P}$ is a symmetric positive definite matrix that satisfies the Lyapunov equation

$$0 = \mathbf{A}^T\mathbf{P} + \mathbf{P}\mathbf{A} + \mathbf{Q} \tag{26}$$

for some symmetric positive definite matrix $\mathbf{Q}$. The matrix $\mathbf{P}$ parameterizes our family of *candidate* Lyapunov certificates for the nonlinear system, which means we would need to find a $\mathbf{P}$ that satisfies

$$\dot{V}(x) = \left.\frac{\partial V(x)}{\partial x}\right|_{x=x} f(x) = x^T\mathbf{P}f(x) + f^T(x)\mathbf{P}x \leq 0$$

for all $x$. This would make $V(x) = x^T\mathbf{P}x$ a Lyapunov function for the "nonlinear" system.

Consider an LTI system

$$\dot{x}(t) = \mathbf{A}x(t)$$

and as suggested in the preceding section, consider a *candidate* Lyapunov function

$$V(x) = x^T\mathbf{P}x$$

where $\mathbf{P}$ is a symmetric positive definite matrix. Because $\mathbf{P} = \mathbf{P}^T > 0$, we already know that $V$ is positive definite. So to show that $V$ is a Lyapunov function for the LTI system we only need to identify conditions under which $\dot{V}$ is negative definite. In particular, we can see that

$$\dot{V}(x) = \frac{\partial V(x)}{\partial x}\mathbf{A}x = 2x^T\mathbf{P}\mathbf{A}x = x^T(\mathbf{A}^T\mathbf{P} + \mathbf{P}\mathbf{A})x$$

This last equality is obtained by recognizing that

$$x^T\mathbf{P}\mathbf{A}x = \left[x^T\mathbf{P}\mathbf{A}x\right]^T = x^T\mathbf{A}^T\mathbf{P}x$$

We write the equation in this way because the matrix in the parentheses $(\mathbf{A}^T\mathbf{P} + \mathbf{P}\mathbf{A})$ is symmetric with real eigenvalues, thereby making it easier to check if $\dot{V}$ is negative definite. In particular this means that $\dot{V}$ is negative definite if and only if the matrix is $\mathbf{A}^T\mathbf{P} + \mathbf{P}\mathbf{A}$ is negative definite, which is so if and only if all of its eigenvalues are negative. So let $\mathbf{Q}$ be any symmetric positive definite matrix, $\mathbf{Q}$. If we can find a symmetric positive definite matrix $\mathbf{P}$ that satisfies the Lyapunov equation

$$\mathbf{A}^T\mathbf{P} + \mathbf{P}\mathbf{A} + \mathbf{Q} = 0$$

then this would imply that

$$\dot{V}(x) = -x^T \mathbf{Q} x < 0$$

for all $x \neq 0$. So $V$ is negative definite and on the basis of Theorem 7 we can conclude that it is a certificate for the asymptotic stability of the origin. Our preceding discussion can now be summarized in the following theorem

THEOREM 8. **(Direct Method for LTI System)** *If there exist symmetric positive definite matrices* $\mathbf{P}$ *and* $\mathbf{Q}$ *such that*

$$\mathbf{A}^T \mathbf{P} + \mathbf{P}\mathbf{A} + \mathbf{Q} = 0$$

*then the origin of* $\dot{x}(t) = \mathbf{A}x(t)$ *is asymptotically stable.*

## 2. Advanced Lyapunov Stability Theory for LTI Systems

This section discusses two advanced stability results relevant to LTI systems. The first result is a *converse theorem* that establishes that the existence of a Lyapunov function is necessary and sufficient for asymptotic stability. The second result is called the *indirect method* and provides a way to estimate the neighborhood about an equilibrium for which linear models capture the qualitatitive behavior of a nonlinear system.

**2.1. Converse LTI Theorem:** Recall that we already established a necessary and sufficient condition for the origin to be asymptotically stable. This was obtained by writing out the solution to the state equation as

$$x(t) = \sum_{i=1}^{\sigma} \sum_{k=1}^{m_i} \mathbf{A}_{ik} t^k e^{\lambda_i t} x_0$$

for $t \geq 0$ where $\mathbf{A}_{ik}$ is the residue matrix, $\sigma$ is the number of distinct eigenvalues of $\mathbf{A}$, $\lambda_i$ is the $i$th distinct eigenvalue of $\mathbf{A}$ with multiplicity $m_i$. In the above equation, one readily sees that $x(t) \to 0$ as $t \to \infty$ if and only if $\mathrm{Re}(\lambda_i) < 0$ for all $i = 1, 2, \ldots, \sigma$. So while Theorem 8 provides a "sufficient" condition for the origin to be asymptotically stable, the eigenvalue condition ($\mathrm{Re}(\lambda_i) < 0$ for all $i$) is a stronger condition in that

it is *necessary and sufficient*. A matrix, $\mathbf{A}$, that satisfies this condition is said to be *Hurwitz*. This result is important enough that we also summarize it as a theorem.

THEOREM 9. *The origin of the LTI system, $\dot{x}(t) = \mathbf{A}x(t)$, is globally asymptotically stable if and only if all eigenvalues of $\mathbf{A}$ have negative real parts.*

Since we have this strong necessary and sufficient condition for stability, is it possible to strengthen Lyapunov's direct method so the existence of a Lyapunov function is also necessary and sufficient for the origin of an LTI system to be asymptotically stable? If so, what advantages (if any) does the Lyapunov analysis provide over the eigenvalue analysis? Such results are called *converse theorems*.

We already know that

eigenvalue condition $\Leftrightarrow$ asymptotic stability $\Leftarrow$ Lyapunov condition

From the above implications, it should be clear that if we could show the eigenvalue condition always implies the existence of a Lyapunov function, then we would obtain our converse result (i.e. asymptotic stability $\Rightarrow$ Lyapunov condition).

So let us assume that $\mathbf{A}$ satisfies the eigenvalue condition (i.e. $\mathrm{Re}(\lambda_i) < 0$ for all $i$) and consider a specific candidate Lyapunov function $V(x) = x^T\mathbf{P}x$ where

$$\mathbf{P} = \int_0^\infty e^{\mathbf{A}^T t}\mathbf{Q}e^{\mathbf{A}t}dt \tag{27}$$

in which $\mathbf{Q}$ is any positive definite symmetric matrix. Note that

$$\begin{aligned}
\mathbf{A}^T\mathbf{P} + \mathbf{P}\mathbf{A} &= \int_0^\infty \left(\mathbf{A}^T e^{\mathbf{A}^T t}\mathbf{Q}e^{\mathbf{A}t} + e^{\mathbf{A}^T t}\mathbf{Q}e^{\mathbf{A}t}\mathbf{A}\right) dt \\
&= \int_0^\infty \frac{d}{dt} e^{\mathbf{A}^T t}\mathbf{Q}e^{\mathbf{A}t}dt \\
&= e^{\mathbf{A}^T t}\mathbf{Q}e^{\mathbf{A}t}\Big|_0^\infty = -\mathbf{Q}
\end{aligned}$$

where we used the fact that $e^{\mathbf{A}0} = \mathbf{I}$ and the fact that $\lim_{t\to\infty} e^{\mathbf{A}t} = 0$ because of the eigenvalue condition. The preceding relationships therefore show that for this choice of $\mathbf{P}$

$$\dot{V}(x) = x^T(\mathbf{A}^T\mathbf{P} + \mathbf{P}\mathbf{A})x = -x^T\mathbf{Q}x$$

which means $\dot{V}$ is negative definite.

To complete the certification of $V$ as a Lyapunov function, we need to verify that $V$ is positive definite. Since $\mathbf{Q}$ is symmetric and positive definite, it can be factored as

$$\mathbf{Q} = \mathbf{M}^T\mathbf{M}.$$

where $M$ is nonsingular. So if we look at $V(x)$ we get

$$\begin{aligned} V(x) &= x^T\mathbf{P}x \\ &= x^T\left[\int_0^\infty e^{\mathbf{A}^T t}\mathbf{Q}e^{\mathbf{A}t}dt\right]x \\ &= \int_0^\infty x^T e^{\mathbf{A}^T t}\mathbf{M}^T\mathbf{M}e^{\mathbf{A}t}x dt \\ &= \int_0^T \left|\mathbf{M}e^{\mathbf{A}\tau}x\right|^2 d\tau \geq 0 \end{aligned}$$

So this shows $V(x)$ is positive *semi*definite. Note, however, that when $V(x) = 0$, the above relation implies

$$\left|\mathbf{M}e^{\mathbf{A}\tau}x\right| = 0$$

which can only happen if $\mathbf{M}e^{\mathbf{A}\tau}x = 0$ for all $\tau$. But, we already know $\mathbf{M}$ and $e^{\mathbf{A}\tau}$ (transition matrix for a continuous-time system) are both non-singular, which means $x = 0$ and so $V(x)$ is positive definite since the only time $V(x) = 0$ is when $x = 0$. What we've just done is prove that when the eigenvalue condition holds, then there is a Lyapunov certificate, $V(x) = x^T\mathbf{P}x$ where $\mathbf{P}$ is given by equation (27). From the direct method, we know that the Lyapunov condition implies asymptotic stability. What we have shown above is that if the origin is asymptotically stable then the

Lyapunov condition must hold. This result is important enough that we formalize it in a theorem

THEOREM 10. *(***Converse Theorem:***) The origin of* $\dot{x}(t) = \mathbf{A}x(t)$ *is asymptotically stable if and only if for* any *positive definite symmetric matrix* $\mathbf{Q}$*, there exists a symmetric positive definite matrix,* $\mathbf{P}$ *that satisfies the Lyapunov equation* $\mathbf{A}^T\mathbf{P} + \mathbf{PA} + \mathbf{Q} = 0$.

We now have two necessary and sufficient conditions for asymptotic stability of the origin of an LTI system. Which one to use? If all one wants is a yes/no declaration of stability, then the eigenvalue test is easier and more stable numerically than trying to solve the Lyapunov equation. The utility of Lyapunov methods lies in their use as design tools. To illustrate this, let us consider the inhomogeneous system

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

and the objective is to find a matrix $\mathbf{K}$ such that if we let $u = \mathbf{K}x$ then the origin of the "controlled" system

$$\dot{x}(t) = (\mathbf{A} + \mathbf{BK})x(t)$$

is asymptotically stable.

One could try to find this $\mathbf{K}$ through the eigenvalue condition. In particular, one would look at $\mathbf{K}$ as a matrix of parameters and then form the characteristic polynomial

$$\det(s\mathbf{I} - \mathbf{A} - \mathbf{BK}) = p(s; \mathbf{K}) = 0$$

We would then solve for the roots of this polynomial *as a function of the parameters* $\mathbf{K}$. In general, this is extremely hard to do because the coefficients of $p(s; \mathbf{K})$ are nonlinear functions of the components of $\mathbf{K}$. This approach can be used for simple low-dimensional systems, but it is not very practical for systems with a large state space.

Alternatively, we can use Theorem 10. In this case, we know that if there exists a $\mathbf{K}$ that stabilizes the origin, then the controlled system

$$\dot{x}(t) = (\mathbf{A} + \mathbf{BK})x(t)$$

has a Lyapunov function, $V(x) = x^T\mathbf{P}x$, where for any choice of $\mathbf{Q} = \mathbf{Q}^T > 0$ we find a matrix $\mathbf{P} = \mathbf{P}^T > 0$ that satisfies the Lyapunov equation

$$(\mathbf{A} + \mathbf{BK})^T\mathbf{P} + \mathbf{P}(\mathbf{A} + \mathbf{BK}) + \mathbf{Q} = 0 \tag{28}$$

This Lyapunov equation has two decision variables we would need to solve for (once $\mathbf{Q}$ has been chosen). These decision variables are $\mathbf{P}$ and the gain matrix $\mathbf{K}$.

At first glance, the Lyapunov equation (28) looks difficult to solve because it is bilinear in $\mathbf{P}$ and $\mathbf{K}$. But we can force this equation to be a linear matrix equation by defining a two new matrix variables

$$\begin{aligned} \mathbf{X} &= \mathbf{P}^{-1} \\ \mathbf{Y} &= \mathbf{KX} \end{aligned}$$

and note that if we pre-multiply and post-multiply equation (28) by $\mathbf{P}^{-1}$, we get

$$\begin{aligned} 0 &= \mathbf{P}^{-1}\left\{(\mathbf{A} + \mathbf{BK})^T\mathbf{P} + \mathbf{P}(\mathbf{A} + \mathbf{BK}) + \mathbf{Q}\right\}\mathbf{P}^{-1} \\ &= \mathbf{P}^{-1}\mathbf{A}^T + \mathbf{AP}^{-1} + \mathbf{P}^{-1}\mathbf{K}^T\mathbf{B}^T + \mathbf{BKP}^{-1} + \mathbf{P}^{-1}\mathbf{QP}^{-1} \\ &= \mathbf{XA}^T + \mathbf{AX} + \mathbf{Y}^T\mathbf{B}^T + \mathbf{BY} + \mathbf{XQX} \end{aligned}$$

Recall that we can choose $\mathbf{Q}$ to be any symmetric positive definite matrix, so we are free to choose $\mathbf{R} := \mathbf{XQX}$ to be any symmetric positive definite matrix. Then we have

$$0 = \mathbf{XA}^T + \mathbf{AX} + \mathbf{Y}^T\mathbf{B}^T + \mathbf{BY} + \mathbf{R} \tag{29}$$

This equation only has two decision variables, $\mathbf{X}$ and $\mathbf{Y}$. But what is important to note here is that equation (29) is *linear* in these variables. So we

can easily solve for $\mathbf{X}$ and $\mathbf{Y}$ once we have selected $\mathbf{R}$ and the controller gain is then

$$\mathbf{K} = \mathbf{Y}\mathbf{X}^{-1}$$

computed directly from solution to the linear matrix equation (29).

What this discussion shows is that while the eigenvalue condition for asymptotic stability is easy to verify, it is more difficult to use as a design tool because it leads to nonlinear design equations. For the LTI systems, the use of Lyapunov methods on the other hand lead to linear equations that are much easier to solve, thereby establishing the value of Lyapunov methods as design tools. Below a simple example is used to illustrate the difference in these two methods for stabilizing an LTI system.

**Example:** Let us consider the following LTI system,

$$\dot{x}(t) = \mathbf{A}x(t) + \mathbf{B}u(t) = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} x(t) + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} u(t)$$

The objective is to determine a state feedback control law, $u = \mathbf{K}x$ that stabilizes the origin of this system. We will try doing this using the eigenvalue condition and Lyapunov methods.

For the eigenvalue condition, we first need to determine the characteristic equation for the controlled system. We let $\mathbf{K} = \begin{bmatrix} k_1 & k_2 & k_3 \end{bmatrix}$ and then note that the characteristic polynomial is

$$\begin{aligned} p(s) &= \det(s\mathbf{I} - \mathbf{A} - \mathbf{B}\mathbf{K}) \\ &= \det\left( \begin{bmatrix} s-1 & -1 & 0 \\ 0 & s-1 & -1 \\ 0 & 0 & s-1 \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ k_1 & k_2 & k_3 \end{bmatrix} \right) \\ &= s^3 - (3+k_3)s^2 + (3 - k_2 + 2k_3)s - (1 + k_1 - k_2 + k_3) \end{aligned}$$

We now need to determine those values of $k_1$, $k_2$, and $k_3$ for which all of the characteristic equation roots have negative real parts. The tool we will use

for this is known as the *Routh-Hurwitz* criterion [Ogata (1970)]. The best way of explaining this criterion is by example. So consider a polynomial of the form,

$$p(s) = a_0 s^n + a_1 s^{n-1} + \cdots + a_{n-1} s + a_n$$

where $a_n \neq 0$. For this polynomial we construct the Routh array

$$\begin{array}{c|ccccc} s^n & a_0 & a_2 & a_4 & a_6 & \cdots \\ s^{n-1} & a_1 & a_3 & a_5 & a_7 & \cdots \\ s^{n-2} & b_1 & b_2 & b_3 & b_4 & \cdots \\ s^{n-3} & c_1 & c_2 & c_3 & c_4 & \cdots \\ \vdots & \vdots & \vdots & & & \\ s^2 & k_1 & k_2 & & & \\ s^1 & \ell_1 & & & & \\ s^0 & m_1 & & & & \end{array}$$

where the first two rows are obtained from the coefficients of polynomial, $p(s)$. The third row of the array is obtained by the negative of the determinant of a matrix formed from $2 \times 2$ blocks of coefficients in the preceding 2 rows divided by the first element of the second row. This means that

$$\begin{aligned} b_1 &= -\frac{1}{a_1}\det\begin{bmatrix} a_0 & a_2 \\ a_1 & a_3 \end{bmatrix} = -\frac{a_0 a_3 - a_1 a_2}{a_1} \\ b_2 &= -\frac{1}{a_1}\det\begin{bmatrix} a_2 & a_4 \\ a_3 & a_5 \end{bmatrix} = -\frac{a_2 a_5 - a_3 a_4}{a_1} \\ &\vdots \end{aligned}$$

The fourth row is computed in the same manner so that

$$\begin{aligned} c_1 &= -\frac{1}{b_1}\det\begin{bmatrix} a_1 & a_3 \\ b_1 & b_2 \end{bmatrix} = -\frac{a_1 b_2 - b_1 a_3}{b_1} \\ c_2 &= -\frac{1}{b_1}\det\begin{bmatrix} a_3 & a_5 \\ b_2 & b_3 \end{bmatrix} = -\frac{a_3 b_3 - b_2 a_5}{b_1} \end{aligned}$$

We continue in this manner until the end of the array. *The number of roots in the open right half plane (i.e. unstable poles) is equal to the number of sign changes in the first column of the completed Routh array.*

So let us return to our characteristic equation,

$$\begin{aligned} 0 &= s^3 - (3+k_3)s^2 + (3-k_2+2k_3)s - (1+k_1-k_2+k_3) \\ &= a_0 s^3 + a_1 s^2 + a_2 s + a_3 \end{aligned}$$

and we have $a_0 = 1$, $a_1 = -3 - k_3$, $a_2 = 3 - k_2 + 2k_3$, and $a_3 = -1 - k_1 + k_2 - k_3$. So the first two rows of the Routh array are

$$\begin{array}{c|cc} s^3 & 1 & 3-k_2+2k_3 \\ s^2 & -3-k_3 & -1-k_1+k_2-k_3 \end{array}$$

We compute the third row as

$$\begin{aligned} b_1 &= -\frac{1}{-3-k_3}\det\begin{bmatrix} 1 & 3-k_2+2k_3 \\ -3-k_3 & -1-k_1+k_2-k_3 \end{bmatrix} \\ &= \frac{-k_1-2k_2+8k_3-k_2k_3+2k_3^2+8}{3+k_3} \\ b_2 &= 0 \end{aligned}$$

which gives us the array

$$\begin{array}{c|cc} s^3 & 1 & 3-k_2+2k_3 \\ s^2 & -3-k_3 & -1-k_1+k_2-k_3 \\ s^1 & \frac{-k_1-2k_2+8k_3-k_2k_3+2k_3^2+8}{3+k_3} & 0 \end{array}$$

The last row of the array is

$$\begin{aligned} c_1 &= -\frac{1}{b_1}\det\begin{bmatrix} a_1 & a_3 \\ b_1 & 0 \end{bmatrix} \\ &= -1-k_1+k_2-k_3 \\ c_2 &= 0 \end{aligned}$$

and the completed Routh array is now

$$\begin{array}{c|cc} s^3 & 1 & 3-k_2+2k_3 \\ s^2 & -3-k_3 & -1-k_1+k_2-k_3 \\ s^1 & \frac{-k_1-2k_2+8k_3-k_2k_3+2k_3^2+8}{3+k_3} & 0 \\ s^0 & -1-k_1+k_2-k_3 & \\ & 0 & \end{array}$$

So to ensure that all roots of the characteristic polynomial have non-positive real parts, we need to select $k_1$, $k_2$, and $k_3$ so there are no sign changes in the first column of the Routh array. The first element of this column is $1$, so that means that

$$-3-k_3 \geq 0 \quad \Rightarrow \boxed{k_3 \leq -3}$$

So let $k_3 = -4$. For this choice of $k_3$, we see that the sign of the third element will be determined by

$$\begin{aligned} 0 &\leq k_1+2k_2-8k_3+k_2k_3-2k_3^2-8 \\ &= \boxed{k_1-2k_2-8} \end{aligned}$$

The sign of the last entry in the first column of the Routh array is determined by

$$0 \leq -1-k_1+k_2-k_3 = \boxed{3-k_1+k_2}$$

So assuming $k_3 = -4$, we need to choose $k_1$ and $k_2$ so

$$\begin{aligned} 0 &\leq 3-k_1+k_2 \\ 0 &\leq -8+k_1-2k_2 \end{aligned}$$

which will be satisfied if we section $k_2 = -7$ and $k_1 = -5$. If we insert these choices back into the original characteristic equation, we obtain

$$p(s) = s^3+s^2+2s+1$$

which has one root at $-0.57$ and two others at $-0.215 \pm 1.30j$. So all real parts are non-positive.

Let us now see how this analysis would have been done using Lyapunov methods. In this case, we form the linear matrix equation

$$0 = \mathbf{X}\mathbf{A}^T + \mathbf{A}\mathbf{X} + \mathbf{Y}^T\mathbf{B}^T + \mathbf{B}\mathbf{Y} + \mathbf{I}$$

where I've chosen $\mathbf{R}$ to be the identity matrix. This must be solved subject to $\mathbf{X} > 0$. This is an example of a linear matrix equation that is easily solved using convex optimization methods used in solving linear matrix inequalities (LMI) [Boyd et al. (1994)]. To do this, we recast our problem as a constrained optimization problem of the form

$$\begin{array}{rll} \text{maximize:} & \text{trace}(\mathbf{X}) & \\ \text{with respect to:} & \mathbf{X} \text{ and } \mathbf{Y} & \\ \text{subject to:} & \epsilon_1\mathbf{I} \leq \mathbf{X} & \\ & \epsilon_2\mathbf{I} \leq -\mathbf{X}\mathbf{A}^T - \mathbf{A}\mathbf{X} - \mathbf{B}\mathbf{Y} - \mathbf{B}^T\mathbf{Y}^T & \end{array}$$

where $\epsilon_1 > 0$ and $\epsilon_2 > 0$ are positive real constants we are free to choose. By maximizing the trace of $\mathbf{X}$, we find a solution that gets close to the boundary of the sets generated by the linear matrix inequalities. The constants $\epsilon_1$ and $\epsilon_2$ are chosen to enforce how "definite" we want these inequalities to be.

The preceding optimization problem can be efficiently solved using convex optimization programs known as semi-definite programming (SDP) solvers [Toh et al. (1999)]. In general, one uses a programming interface to access these SDP solvers. The one I'll use below is YALMIP [Lofberg (2004)]. For this example I used the the following script to solve for the controller gains,

```
clear all;
%declare system matrices
A = [1 1 0; 0 1 1; 0 0 1];
B = [0;0;1];
n  = size(A,1);
%declare SDP variables X and Y
X  = sdpvar(n,n);
Y  = sdpvar(1,n);

eps1 =1;  eps2 = 1;
```

```
%declare LMI's
F = [X > eps1*eye(n)];
F = [F,-X*A'-A*X-B*Y-Y'*B'-eps2*eye(n)];
%solve the SDP that maximizes trace of X subject to F
optimize(F, trace(X))

P = inv(value(X));
K = value(Y)*P

eig(A+B*K)
```

The output generated by this script is shown below

```
>> test_file

 num. of constraints =  9
 dim. of sdp    var  =  6,   num. of sdp  blk  =  2
*******************************************************************
   SDPT3: Infeasible path-following algorithms
*******************************************************************
 version  predcorr  gam  expon  scale_data
   HKM      1      0.000   1        0
it pstep dstep pinfeas dinfeas  gap      prim-obj      dual-obj    cputime
-------------------------------------------------------------------
 0|0.000|0.000|1.4e+01|7.8e+00|6.0e+02|-6.000000e+01  0.000000e+00| 0:0:00| chol  1  1
 1|0.824|0.807|2.5e+00|1.6e+00|1.1e+02|-3.895285e+01 -1.384793e+01| 0:0:00| chol  1  1
 2|0.833|0.545|4.3e-01|7.1e-01|3.4e+01|-6.307726e+01 -2.301790e+01| 0:0:00| chol  1  1
 3|0.610|1.000|1.7e-01|7.1e-04|5.9e+01|-5.093197e+01 -9.675923e+01| 0:0:00| chol  1  1
 4|0.970|0.916|5.0e-03|1.2e-04|5.6e+00|-6.416986e+01 -6.946985e+01| 0:0:00| chol  1  1
 5|0.974|0.984|1.3e-04|1.4e-04|1.4e-01|-6.734862e+01 -6.746674e+01| 0:0:00| chol  1  1
 6|0.919|0.983|1.1e-05|1.1e-05|8.1e-03|-6.740433e+01 -6.741073e+01| 0:0:00| chol  1  2
 7|0.927|0.991|9.8e-07|4.7e-07|3.9e-04|-6.740698e+01 -6.740727e+01| 0:0:00| chol  2  2
 8|0.952|0.983|2.4e-07|2.1e-08|1.3e-05|-6.740708e+01 -6.740710e+01| 0:0:00| chol  2  2
 9|0.972|0.985|2.9e-08|6.1e-10|3.1e-07|-6.740709e+01 -6.740709e+01| 0:0:00|
  stop: max(relative gap, infeasibilities) < 1.00e-07
-------------------------------------------------------------------
 number of iterations   =  9
 primal objective value = -6.74070920e+01
 dual   objective value = -6.74070931e+01
 gap := trace(XZ)       = 3.06e-07
 relative gap           = 2.25e-09
 actual relative gap    = 8.09e-09
 rel. primal infeas     = 2.88e-08
 rel. dual   infeas     = 6.09e-10
```

```
 norm(X), norm(y), norm(Z) = 5.2e+01, 8.3e+01, 5.9e+01
 norm(A), norm(b), norm(C) = 9.3e+00, 2.7e+00, 3.4e+00
 Total CPU time (secs)  = 0.34
 CPU time per iteration = 0.04
 termination code       =  0
 DIMACS: 3.9e-08  0.0e+00  1.1e-09  0.0e+00  8.1e-09  2.3e-09
------------------------------------------------------------------

ans =

  struct with fields:

    yalmiptime: 1.1079
    solvertime: 0.5257
          info: 'Successfully solved (SDPT3-4)'
       problem: 0


P =

    0.5937    0.4511    0.0937
    0.4511    0.4538    0.1016
    0.0937    0.1016    0.0441


K =

  -10.5098  -12.2276   -4.0915


ans =

  -0.2001 + 0.0000i
  -0.4457 + 2.5822i
  -0.4457 - 2.5822i

>>
```

The first part of the output is generated by the SDP solver in which the termination code of $0$ indicates the SDP was successfully solved. The last

part shows the resulting $\mathbf{P}$ and the control gains,

$$\mathbf{K} = \begin{bmatrix} -10.5 & -12.2 & -4.09 \end{bmatrix}$$

The closed loop eigenvalues are seen to be $-0.2$ and $-0.44 \pm 2.58j$, which indeed shows that we've stabilized the equilibrium. Notice that solving the problem in this manner was tremendously easy because it could be put in a linear form for which existing optimization tools could be used. The eigenvalue approach provided a systematic way to identify stabilizing controller gains, but the Routh-Hurwitz array, in general, creates a set of nonlinear multivariate polynomial constraints that can be extremely difficult to solve by hand or by the computer. We therefore see the value in the Lyapunov method lies in how it formulates the stabilization problem as a convex problem that allows one to easily use efficient codes to solve it.

**2.2. Indirect Method:** We can use Lyapunov stability concept to establish when the linearized system's stability can be used to infer the stability of the nonlinear system. This theorem is called *Lyapunov's indirect method* [Khalil (2002)].

THEOREM 11. **Lyapunov's Indirect Method:** *Let $\dot{x}(t) = \mathbf{A}x(t)$ be the linearization of a nonlinear $\dot{x}(t) = f(x)(t)$ about an equilibrium at the origin. Let $\{\lambda_i\}_{i=1}^n$ denote the eigenvalues of $\mathbf{A}$. If $\mathrm{Re}(\lambda_i) < 0$ for all $i = 1, 2, \ldots, n$, then the origin of the nonlinear system is asymptotically stable. If $\mathrm{Re}(\lambda_i) > 0$ for any $i \in \{1, 2, \ldots, n\}$, then the origin of the nonlinear system is unstable.*

**Proof:** The "stability" part of the theorem can be proven using what we already have. The "instability" part of the theorem is proven using a instability certificate known as a Chetaev function. These instability certificates are not of direct interest in this course, so we will only prove the stability part of the theorem.

Consider a candidate Lyapunov function of the form $V(x) = x^T \mathbf{P} x$ where $\mathbf{P} = \mathbf{P}^T > 0$. Under the first condition when $\mathbf{A}$ is Hurwitz, we can

take $\mathbf{P}$ to satisfy the Lyapunov equation $\mathbf{A}^T\mathbf{P} + \mathbf{P}\mathbf{A} + \mathbf{Q} = 0$ for some $\mathbf{Q} = \mathbf{Q}^T > 0$. Take the directional derivative of $V$ with respect to the nonlinear system's vector field, $f$, and get

$$\begin{aligned}
\dot{V} &= x^T\mathbf{P}f(x) + f^T(x)\mathbf{P}x \\
&= x^T\mathbf{P}(\mathbf{A}x + g(x)) + \left[x^T\mathbf{A}^T + g^T(x)\right]\mathbf{P}x \\
&= x^T(\mathbf{P}\mathbf{A} + \mathbf{A}^T\mathbf{P})x + 2x^T\mathbf{P}g(x) \\
&= -x^T\mathbf{Q}x + 2x^T\mathbf{P}g(x)
\end{aligned}$$

So the first term is negative definite. The second term is indefinite. We know, however, that $\frac{|g(x)|}{|x|} \to 0$ as $|x| \to 0$ so for any $\gamma > 0$ there exists $r > 0$ such that

$$|g(x)| < \gamma|x|$$

when $|x| < r$. This means that

$$\begin{aligned}
\dot{V} &< -x^T\mathbf{Q}x + 2\gamma\|\mathbf{P}\||x|^2 \\
&< -(\underline{\lambda}(\mathbf{Q}) - 2\gamma\|\mathbf{P}\|)|x|^2
\end{aligned}$$

where $\underline{\lambda}(\mathbf{Q})$ is the minimum eigenvalue of $\mathbf{Q}$. So if we choose $\gamma < \frac{\underline{\lambda}(\mathbf{Q})}{2\|\mathbf{P}\|}$ then we can guarantee $\dot{V} < 0$ for $|x| < r$ which establishes the asymptotic stability of the origin when $\mathbf{A}$ is Hurwitz. $\diamondsuit$

We summarize the preceding theorem's findings below

- If the equilibrium of the linearization of $\dot{x} = f(x)$ is asymptotically stable, then the origin of the nonlinear system is also locally asymptotically stable.
- If any eigenvalue of the linearization has a positive real part then the origin of the nonlinear system is unstable.
- If all eigenvalues of the linearization have nonpositive real parts and there exists at least one eigenvalue with a zero real part, then nothing can be concluded about the stability of the equilibrium.

The theorem says that if the origin of a nonlinear system's linearization is asymptotically stable, then so too is the origin of the original system provided the equilibrium is hyperbolic (the linearization's eigenvalues have no zero real parts). A similar finding holds for the certain unstable linearizations. This means, therefore, that stabilizing the linearization is often good enough to stabilize a real-life nonlinear plant. The only time it cannot be used is when the none of the algorithms have positive real parts and at least one has a zero real part. This result, of course, only holds in a neighborhood of the origin and that neighborhood may be too small for practical use depending on the system under study.

**Example:** Let us apply this to a nonlinear pendulum system,

$$\begin{aligned}\dot{x}_1(t) &= x_2(t)\\ \dot{x}_2(t) &= -\frac{g}{\ell}\sin x_1(t) - x_2(t)\end{aligned}$$

where $x_1$ is the pendulum angle, $x_2$ is the angle's time derivative, $g$ is gravitational acceleration, $\ell$ is the length of the pendulum, and the mass of the pendulum bob is $1$. We will examine the stability of the physical equilibria at $(0,0)$ and $(\pi,0)$ using the indirect method. This requires that we first compute the Jacobian of

$$f(x_1,x_2) = \begin{bmatrix} f_1(x_1,x_2) \\ f_2(x_1,x_2) \end{bmatrix} = \begin{bmatrix} x_2 \\ -\frac{g}{\ell}\sin x_1 - x_2 \end{bmatrix}$$

That Jacobian is

$$\left[\frac{\partial f}{\partial x}\right] = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\frac{g}{\ell}\cos x_1 & -1 \end{bmatrix}$$

At the two equilibria, $(0,0)$, and $(\pi,0)$ we end up with the matrices,

$$\begin{aligned}\mathbf{A}_1 &= \left[\frac{\partial f}{\partial x}\right]_{x=(0,0)} = \begin{bmatrix} 0 & 1 \\ -\frac{g}{\ell} & -1 \end{bmatrix}\\ \mathbf{A}_2 &= \left[\frac{\partial f}{\partial x}\right]_{x=(\pi,0)} = \begin{bmatrix} 0 & 1 \\ \frac{g}{\ell} & -1 \end{bmatrix}\end{aligned}$$

The characteristic polynomial of $\mathbf{A}_1$ is $s(s+1)+\frac{g}{\ell}=0$ which as roots

$$\lambda_{1,2}=-\frac{1}{2}\pm\frac{1}{2}\sqrt{1-4\frac{g}{\ell}}$$

both roots have negative real parts for all $g/\ell>0$ and so the indirect method implies the equilibrium at $(0,0)$ is asymptotically stable. The characteristic polynomial for $\mathbf{A}_2$ has roots

$$\lambda_{1,2}=-\frac{1}{2}\pm\frac{1}{2}\sqrt{1+4\frac{g}{\ell}}$$

One root has a positive real part and the other has a negative real part. There are no center (zero real part) eigenvalues so by the theorem we know the equilibrium at $(\pi,0)$ is unstable.

## 3. Lyapunov Stability for Discrete-time LTI Systems:

Consider the homogeneous time-invariant system

$$x(k+1)=f(x(k))$$

$x^*\in\mathbb{R}$ is an equilibrium point of $x^*=f(x^*)$ and without a loss of generality we assume $x^*=0$. We introduce a candidate Lyapunov function $V:\mathbb{R}^n\to\mathbb{R}$ which is positive definite and define the first difference of $V$ with respect to a given state trajectory as

$$\Delta V(x)=V(f(x))-V(x)$$

Our usual notions of Lyapunov stability are

- The equilibrium is stable (Lyapunov) if for all $\epsilon>0$ there exists $\delta>0$ such that $|x(0)|<\delta$ implies $|x(k)|<\epsilon$ for all $k\geq 0$.
- The system is unstable if it is not stable.
- The equilibrium is asymptotically stable if the origin is stable and for $|x_0|$ small enough, we can show $x(k)\to 0$ as $k\to\infty$.

Lyapunov's direct method now relies on using first differences. In particular this means

- The origin is stable if $V > 0$ and $\Delta V \leq 0$
- The origin is asymptotically stable if $V > 0$ and $\Delta V < 0$.

If we confine our attention to LTI systems of the form

$$x(k+1) = \mathbf{A}x(k)$$

and consider a candidate Lyapunov function of the form $V(x) = x^T\mathbf{P}x$ where $\mathbf{P} = \mathbf{P}^T > 0$ then the first difference condition becomes

$$\begin{aligned} \Delta V(x)(k) &= V(\mathbf{A}x(k)) - V(x(k)) \\ &= x^T(k)(\mathbf{A}^T\mathbf{P}\mathbf{A} - \mathbf{P})x(k) \leq 0 \end{aligned}$$

which leads to the discrete-time Lyapunov equation

$$\mathbf{A}^T\mathbf{P}\mathbf{A} - \mathbf{P} + \mathbf{Q} = 0 \tag{30}$$

where $\mathbf{P}$ and $\mathbf{Q}$ are symmetric positive definite matrices. In particular for any $\mathbf{Q} = \mathbf{Q}^T > 0$, if we can find a $\mathbf{P} = \mathbf{P}^T > 0$ that satisfies the discrete time Lyapunov equation (30), then the origin of the discrete time system is asymptotically stable. Just as in the continuous-time case, the existence of such a Lyapunov function is necessary and sufficient for asymptotic stability and is therefore equivalent to the eigenvalue condition that all eigenvalues lie in the unit circle of the complex plane.

While one may use Lyapunov analysis to certify whether the origin of a discrete-time LTI system is asymptotically stable, it is actually easier to use the eigenvalues of $\mathbf{A}$ to make this assessment. In particular, for the LTI system

$$x(k+1) = \mathbf{A}x(k)$$

Let $\lambda_1, \ldots, \lambda_\sigma$ denote the $\sigma$ distinct eigenvalues of $\mathbf{A}$ and let $m_i$ denote the algebraic multiplicity of the $i$th eigenvalue. The state transition matrix becomes

$$\mathbf{\Phi}(k;0) = \mathbf{A}^k \quad = \quad \sum_{i=1}^{\sigma} \left[ \mathbf{A}_{i0} u(k) + \sum_{\ell=1}^{m_i - 1} \mathbf{A}_{i\ell} k(k-1)\ldots(k-\ell+1) \lambda_i^{k-\ell} u(k-\ell) \right]$$

where

$$\mathbf{A}_{i\ell} = \frac{1}{\ell!} \lim_{z \to \lambda_k} \left\{ \frac{d^{m_i - \ell - 1}}{dz^{m_i - \ell - 1}} \left[ (z - \lambda_i)^{m_i} (z\mathbf{I} - \mathbf{A})^{-1} \right] \right\}$$

with $u(k)$ denoting the unit step function. Note that all terms in the preceding expression asymptotically go to zero as $k \to \infty$ if and only if $|\lambda_i| < 1$ for all $i = 1, \ldots, \sigma$. So we can say the origin of the discrete-time LTI system is asymptotically stable if and only if all eigenvalues of $\mathbf{A}$ lie within the unit circle (i.e. $|\lambda_i| < 1$ for all $i = 1, 2 \ldots, \sigma$).

Just as we had the Routh-Hurwitz criterion to assess a continuous-time LTI system's stability, we can use the *Jury Stability Test* [Ogata et al. (1995)] to do the same for discrete-time LTI systems. In this case, let us assume that the discrete-time system's characteristic equation can be written as

$$d(z) = a_n z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0 = 0$$

with $a_n > 0$. Jury's stability test is applied to the following array

| $z^0$ | $z^1$ | $z^2$ | $\cdots$ | $z^{n-k}$ | $\cdots$ | $z^{n-1}$ | $z^n$ |
|---|---|---|---|---|---|---|---|
| $a_0$ | $a_1$ | $a_2$ | $\cdots$ | $a_{n-k}$ | $\cdots$ | $a_{n-1}$ | $a_n$ |
| $a_n$ | $a_{n-1}$ | $a_{n-2}$ | $\cdots$ | $a_k$ | $\cdots$ | $a_1$ | $a_0$ |
| $b_0$ | $b_1$ | $b_2$ | $\cdots$ | $b_{n-k}$ | $\cdots$ | $b_{n-1}$ | |
| $b_{n-1}$ | $b_{n-2}$ | $b_{n-3}$ | $\cdots$ | $b_{k-1}$ | $\cdots$ | $b_0$ | |
| $c_0$ | $c_1$ | $c_2$ | $\cdots$ | $c_{n-k}$ | $\cdots$ | | |
| $c_{n-2}$ | $c_{n-3}$ | $c_{n-4}$ | $\cdots$ | $c_{k-2}$ | $\cdots$ | | |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | | | |
| $\ell_0$ | $\ell_1$ | $\ell_2$ | $\ell_3$ | | | | |
| $\ell_3$ | $\ell_2$ | $\ell_1$ | $\ell_0$ | | | | |
| $m_0$ | $m_1$ | $m_2$ | | | | | |

where

$$b_k = \det \begin{bmatrix} a_0 & a_{n-k} \\ a_n & a_k \end{bmatrix}, \quad c_k = \det \begin{bmatrix} b_0 & b_{n-1-k} \\ b_{n-1} & b_k \end{bmatrix}, \quad d_k = \det \begin{bmatrix} c_0 & c_{n-2-k} \\ c_{n-2} & c_k \end{bmatrix}, \ldots$$

The necessary and sufficient conditions for $d(z)$ to have no roots on or outside the unit circle with $a_n > 0$ are

$$\begin{aligned} d(1) &> 0 \\ (-1)^n d(-1) &> 0 \\ |a_0| &< a_n \\ |b_0| &> |b_{n-1}| \\ |c_0| &> |c_{n-2}| \\ |d_0| &> |d_{n-3}| \\ &\vdots \\ |m_0| &> |m_2| \end{aligned}$$

As an example, let us consider a discrete-time system whose characteristic polynomial is

$$d(z) = z^3 - 1.8z^2 + 1.05z - 0.20 = 0$$

The first conditions of the Jury test are

$$\begin{aligned} d(1) &= 1 - 1.8 + 1.05 - 0.2 = 0.05 > 0 \\ (-1)^3 d(-1) &= -[-1 - 1.8 - 1.05 - 0.2] = 4.05 > 0 \\ |a_0| &= 0.2 < a_3 = 1 \end{aligned}$$

The Jury array is

| $z^0$ | $z^1$ | $z^2$ | $z^3$ |
|---|---|---|---|
| $-0.2$ | $1.05$ | $-1.8$ | $1$ |
| $1$ | $-1.8$ | $1.05$ | $-0.2$ |
| $-0.96$ | $1.59$ | $-0.69$ | |

So the last condition is

$$|b_0| = 0.96 > |b_2| = 0.69$$

Since all conditions are satisfied, all roots of the characteristic equation lie within the unit circle. In particular, one may factor the characteristic equation as

$$d(z) = (z - 0.5)^2(z - 0.8)$$

which is consistent with what was predicted by the Jury test.

An "indirect" result can also be established for discrete time systems where the nonlinear system is written as

$$x(k+1) = \mathbf{A}x(k) + g(x(k))$$

and $g$ is a little-$o$ function. Then if all eigenvalues of $\mathbf{A}$ are in the unit circle the origin of the nonlinear system is asymptotically stable. If at least one eigenvalue of $\mathbf{A}$ is outside the unit circle, then the origin of the nonlinear system is unstable.

## 4. Uniform Stability Concepts

We need to *refine* the earlier Lyapunov stability concept when it is applied to time-varying systems. To help illustrate the need for this refinement, consider the LTV system

$$\dot{x}(t) = (6t\sin(t) - 2t)x(t)$$

One may view this as a linear system, $\dot{x}(t) = a(t)x(t)$ whose time-varying coefficient $a(t)$ is subject to a damping force and a sinusoidal perturbation that both grow over time. The solution may be obtained by first separating the variables, $x$ and $t$

$$\frac{dx}{x} = (6t\sin(t) - 2t)dt$$

and then integrating both sides of the equation from $t_0$ (initial time) to $t > t_0$ assuming $x(t_0) = x_0$.

$$\begin{aligned} x(t) &= x(t_0)\exp\left\{\int_{t_0}^{t}(6s\sin(s) - 2s)ds\right\} \\ &= x(t_0)\exp\left\{6\sin(t) - 6t\cos(t) - t^2 - 6\sin(t_0) + 6t_0\cos(t_0) + t_0^2\right\} \end{aligned}$$

For a fixed initial time, $t_0$, one sees that eventually the quadratic term, $t^2$, will denominate the exponential function's behavior, thereby implying that the exponential function in the above equation is bounded above by a function of $t_0$, say $c(t_0)$. It therefore follows that

$$|x(t)| \leq c(t_0)|x(t_0)|, \quad \text{for all } t \geq t_0$$

If we then consider any $\epsilon > 0$ and select the initial neighborhood $\delta(\epsilon) = \dfrac{\epsilon}{c(t_0)}$ then clearly $|x(t)|$ remains with a $\epsilon$-neighborhood of the equilibrium and so we can conclude the equilibrium at $0$ is stable in the sense of Lyapunov.

The issue we face here, however, is that $\delta$ is not just a function of $\epsilon$, it is also a function of the initial time $t_0$. Our worry is that as $t_0 \to \infty$ that $\delta(\epsilon, t_0)$ could approach a finite limit that is, in fact, zero. If we think of $t_0$ as the system's current "age". Then this says as the system ages (i.e. $t_0$ gets larger) our ability to keep $|x(t)| < \epsilon$ is harder and harder as we have to start in a smaller and smaller distance, $\delta$, from the origin.

This is precisely the case with this particular system. Consider a sequence of initial times $\{t_{0n}\}_{n=0}^{\infty}$ where $t_{0n} = 2n\pi$ for $n = 0, 1, 2, \ldots, \infty$. Let us evaluate $x(t)$ exactly $\pi$ time units after $t_{0n}$ to see that

$$\begin{aligned} x(t_{0n} + \pi) &= x(t_{0n})\exp\left\{6(2n+1)\pi - (2n+1)^2\pi^2 + 6(2n\pi) + (2n)^2\pi^2\right\} \\ &= x(t_{0n})\exp\left\{24n\pi + 6\pi - 4n\pi^2 - \pi^2\right\} \\ &= x(t_{0n})\exp\left\{(4n+1)(6-\pi)\pi\right\} \end{aligned}$$

and for any $x(t_0) \neq 0$ we then can see that the ratio

$$\frac{x(t_{0n} + \pi)}{x(t_{0n})} = e^{(6-\pi)\pi(4n+1)} = 7942.2e^{35.918n} \to \infty$$

as $n \to \infty$. In other words, as the system ages in the sense that $t_{0n}$ goes to infinity, we see that ratio of $x$ at $t_{0n} + \pi$ and $t_{0n}$ become unbounded. This means that the origin becomes "less" stable as $t_0 \to \infty$ since $|x(t)|$ gets larger and larger as $t_0 \to \infty$ assuming $x(t_0)$ starts in the same $\delta$-sized neighborhood of the origin.

These issues also become relevant when we consider asymptotic stability of the equilibrium. To be asymptotically stable, we require the origin to be stable and that $x(t) \to 0$ as $t \to \infty$. Formally, this asymptotic behavior may be seen as requiring for any $\epsilon > 0$ there exists a time $T > 0$ and initial neighborhood, $N_\delta(0)$, such that starting $x_0 \in N_\delta(0)$ implies $|x(t)| < \epsilon$ for all $t \geq T$. $T$ represents the time it takes for the system state to reach the desired $\epsilon$-neighborhood. In general this convergence time is a function of $\epsilon$ as well. But if the system is time-varying then we can also expect $T$ to be a function of the initial time $t_0$. Our worry is that as $t_0 \to \infty$ (i.e. as the system ages) we have $T(\epsilon, t_0) \to \infty$. In other words, as the system ages its convergence time gets slower and slower.

We will use the following LTV system

$$\dot{x}(t) = -\frac{x(t)}{1+t}$$

to illustrate this other convergence issue. Again we separate the variables, $x$ and $t$, and integrate from $t_0$ to $t$ to obtain

$$x(t) = x(t_0)\frac{1+t_0}{1+t}$$

The origin is Lyapunov stable since for any $t_0$ we have $|x(t)| \leq |x(t_0)|$ for $t \geq t_0$. So for any $\epsilon > 0$, we can choose $\delta$ so it is independent of $t_0$. Note that the origin, however, is also *asymptotically stable*. So for all $\epsilon$, we can find $T > 0$ such that $|x(t)| < \epsilon$ for all $t \geq t_0 + T$. In particular, given $\epsilon$ we can bound $|x(t)|$ as

$$|x(t)| \leq |x(t_0)|\frac{1+t_0}{1+t_0+T} < \epsilon$$

which can be rearranged to isolate $T$ and get

$$T > t_0 \left( \frac{|x(t_0)|}{\epsilon} - 1 \right) - 1$$

Note that this is a lower bound on the time, $T$, it takes to reach the target $\epsilon$-neighborhood. So $1/T$ may be taken as the "rate" at which the state converges to the origin. But the lower bound on $T$ in the above equation is also a function of $t_0$ and we can readily see that $\frac{1}{T} \to 0$ as $t_0 \to \infty$. In other words, as the system ages (i.e. $t_0$ gets larger), the system's convergence rate, $1/T$, gets slower and slower. In terms of behavior, this system would appear to "stall out" on its approach to the origin.

The preceding concerns motivate a refinement of our earlier Lyapunov stability concept. We say the equilibrium at the origin is *uniformly stable* if for all $\epsilon > 0$ there exists $\delta > 0$ *that is independent of* $t_0$ such that

$$|x(t_0)| \leq \delta, \quad \Rightarrow \quad |x(t)| < \epsilon \quad \text{for all } t \geq t_0.$$

The equilibrium is *uniformly asymptotically stable* if it is uniformly stable and there exists $\delta$ independent of $t_0$ such that for all $\epsilon > 0$ there exists $T > 0$ that is also independent of $t_0$ such that $|x(t)| < \epsilon$ for all $t \geq t_0 + T(\epsilon)$ and all $|x(t_0)| \leq \delta(\epsilon)$. Finally, we say the origin is *globally uniformly asymptotically stable* (GUAS) if it is uniformly stable and $\delta$ can be chosen so that $\delta(\epsilon) \to \infty$ as $\epsilon \to \infty$ and there exist constants $T$ and $\delta$, both independent of $t_0$, such that for any $\epsilon > 0$ we have $|x(t)| < \epsilon$ for all $t \geq t_0 + T(\epsilon)$ when $|x(t_0)| < \delta(\epsilon)$.

**Example:** Consider the LTV system

$$\dot{x}(t) = \begin{bmatrix} -1 & t \\ 0 & -1 \end{bmatrix} x(t)$$

- Use the definition of asymptotic stability to show the origin is asymptotically stable.
- Use the definition of uniform asymptotic stability to show the origin is not UAS.

We take the initial time to be $t_0$ and let $x_{10} = x_1(t_0)$ and $x_{20} = x_2(t_0)$. Note that $x_2(t) = e^{-(t-t_0)}x_{20}$ for all $t \geq t_0$. This means that the first ODE is

$$\dot{x}_1(t) = -x_1(t) + te^{-(t-t_0)}u(t-t_0)$$

where $u$ is a unit step function The solution for this ODE is

$$\begin{aligned} x_1(t) &= e^{-(t-t_0)}x_{10} + \int_{t_0}^{t} \tau e^{-(\tau-t_0)}e^{-(t-\tau)}x_{20}d\tau \\ &= e^{-(t-t_0)}x_{10} + x_{20}e^{-(t-t_0)}\int_{t_0}^{t} \tau d\tau \\ &= e^{-(t-t_0)}\left(x_{10} + \frac{x_{20}}{2}(t^2 - t_0^2)\right) \end{aligned}$$

Note that for $t \geq t_0$ we get

$$\begin{aligned} |x(t)|^2 &= x_1^2(t) + x_2^2(t) \\ &= e^{-2(t-t_0)}\left(x_{20}^2 + x_{10}^2 + x_{10}x_{20}(t^2 - t_0^2) + \frac{x_{20}^2}{4}(t^2 - t_0^2)^2\right) \\ &\leq e^{-2(t-t_0)}\left(|x_0|^2 + |x_0|^2(t^2 - t_0^2) + \frac{|x_0|^2}{4}(t^2 - t_0^2)^2\right) \\ &= |x_0|^2 e^{-2(t-t_0)}\left(1 + (t^2 - t_0^2) + \frac{1}{4}(t^2 - t_0^2)^2\right) \end{aligned}$$

To assess stability, let $\epsilon > 0$ and $t_0 = 0$, then we get

$$|x(t)|^2 \leq |x_0|^2(1 + K)$$

where $K = \max_t e^{-2t}(t^2 + t^4/4)$. We choose $\delta < \dfrac{\epsilon}{\sqrt{1+K}}$ to show the origin is stable. We can also see that

$$|x(t)|^2 \leq |x_0|^2(e^{-2t} + t^2e^{-2t} + t^4e^{-2t}/4) \to 0$$

as $t \to \infty$ which establishes the origin is asymptotically stable.

To assess uniform asymptotic stability, we need to keep $t_0$. In particular, we see this means

$$|x(t)|^2 \leq |x_0|^2 e^{-2(t-t_0)}\left(1 + (t^2 - t_0^2) + \frac{1}{4}(t^2 - t_0^2)\right)$$

The problem is that

$$t^2 - t_0^2 = (t - t_0)^2 + 2t_0(t_0 - t)$$

which means $\delta$ is also a function of $t_0$, not just $t - t_0$. As a result we cannot conclude the origin is uniformly stable and hence cannot be UAS.

## 5. Lyapunov Stability for Linear Time-varying Systems

As before, there is a Lyapunov theorem (direct method) for time-varying systems that certifies the uniform asymptotic stability of the equilibrium. This theorem is stated below without proof since it uses techniques that are not of direct interest to this course. We will use this theorem in establishing uniform stability results for LTV systems.

THEOREM 12. **(Direct Method for Time-Varying Systems)** *Let $x = 0$ be an equilibrium for $\dot{x}(t) = f(t, x)$, and let $V : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$ be $C^1$ in both arguments. If there exist continuous positive definite functions $\underline{W}$, $\overline{W}$, and $W$ all mapping $\mathbb{R}^n$ onto $\mathbb{R}$ such that*

$$\underline{W}(x) \leq V(t, x) \leq \overline{W}(x)$$
$$\frac{\partial V}{\partial t} + \frac{\partial V}{\partial x}(f(t, x) \leq -W(x)$$

*for all $t \geq 0$ and $x \in D$, then the origin is uniformly asymptotically stable.*

Let us compare Theorem 12 to our earlier direct method for time invariant systems. We first see that the Lyapunov function is now a function of time, $t$, and state, $x$, whereas the time-invariant $V$ was only a function of state. This means that to establish similar conditions we need to "bound" the time variation in $V$. So the condition for $V(x) > 0$ now becomes one where $V(t, x)$ is "sandwiched" between two positive definite functions $\underline{W}(x)$ and $\overline{W}(x)$ which are *independent of* $t$. In a similar spirit, we now require

$$\dot{V} = \frac{\partial V}{\partial t} + \frac{\partial V}{\partial x} f(t, x)$$

to be bounded above by a negative definite $-W(x)$ which is also *independent of* $t$. So the conditions in this theorem are essentially the same as those in the direct method for time-invariant systems. Uniform asymptotic stability of the equilibrium is certified when $V(t,x)$ is sandwiched between two positive definite functions of state and $\dot{V}(t,x)$ is bounded above by a negative definite function of state. As before a function $V(t,x)$ that satisfies these conditions is called a Lyapunov function for UAS.

Let us now apply theorem 12 to an LTV system. So consider the LTV system

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

which has an equilibrium at the origin. We let $\mathbf{A}(t)$ be a continuous function of $t$ and suppose there is a $C^1$ symmetric positive definite matrix-valued function $\mathbf{P} : \mathbb{R} \to \mathbb{R}^{n\times n}$ with positive constants $c_1$ and $c_2$ such that

$$0 < c_1\mathbf{I} \leq \mathbf{P}(t) \leq c_2\mathbf{I}, \quad \text{for all } t \geq t_0 \tag{31}$$

and such that $\mathbf{P}(t)$ satisfies the matrix differential equation

$$-\dot{\mathbf{P}}(t) = \mathbf{P}(t)\mathbf{A}(t) + \mathbf{A}^T(t)\mathbf{P}(t) + \mathbf{Q}(t) \tag{32}$$

where $\mathbf{Q}(t)$ is a continuous symmetric positive definite matrix valued function and positive constant $c_3$ such that

$$\mathbf{Q}(t) \geq c_3\mathbf{I} > 0, \quad \text{for all } t \geq t_0 \tag{33}$$

We want to show that $V : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$ taking values

$$V(t,x) = x^T\mathbf{P}(t)x$$

is a Lyapunov function for the LTV system.

Certifying $V$ as a Lyapunov function simply means checking the conditions in Theorem 12. Clearly

$$c_1|x|^2 \leq V(t,x) = x^T\mathbf{P}(t)x \leq c_2|x|^2$$

based on our assumptions on $\mathbf{P}(t)$ in equation (31). So we can take $\underline{W}(x) = c_1|x|^2$ and $\overline{W}(x) = c_2|x|^2$, both of which are clearly positive definite. We

now check the other condition by computing the directional derivative of $V(t,x)$,

$$\begin{aligned}\dot{V}(t,x) &= x^T\dot{\mathbf{P}}(t)x + x^T\mathbf{P}(t)\dot{x} + \dot{x}^T\mathbf{P}(t)x \\ &= x^T\left\{\dot{\mathbf{P}}(t) + \mathbf{P}(t)\mathbf{A}(t) + \mathbf{A}^T(t)\mathbf{P}(t)\right\}x \\ &= -x^T\mathbf{Q}(t)x \leq -c_3|x|^2\end{aligned}$$

where the last line comes from our assumption on $\mathbf{Q}(t)$ in equation (33). So we can take $W(x) = c_3|x|^2$ which is also clearly positive definite. As both conditions of Theorem 12 are satisfied we can conclude that the equilibrium of the LTV system is UAS provided the conditions in equations (31), (32), and (33) are all satisfied.

The Lyapunov conditions in equations (31-32) are only sufficient UAS of an LTV system. Unlike the LTI case, we cannot use eigenvalues as a necessary and sufficient condition for UAS because these eigenvalues are changing over time. The following theorem provides an alternative "eigenvalue" condition for GUAS of the LTV system,

THEOREM 13. *The origin of $\dot{x}(t) = \mathbf{A}(t)x(t)$ with initial condition $x(t_0) = x_0$ is uniformly asymptotically stable (UAS) if and only if there are positive constants, $k$ and $\lambda$, such that the system's state transition matrix satisfies*

$$\|\mathbf{\Phi}(t;t_0)\| \leq ke^{-\lambda(t-t_0)}$$

*for all $t \geq t_0 > 0$.*

**Proof:** Since $\mathbf{\Phi}$ is the system's transition matrix, we have for $t \geq t_0$ that

$$\begin{aligned}|x(t)| &\leq |\mathbf{\Phi}(t;t_0)x(t_0)| \\ &\leq \|\mathbf{\Phi}(t;t_0)\|\ |x(t_0)| \\ &\leq k|x(t_0)|e^{-\lambda(t-t_0)}\end{aligned}$$

Since this is true for any $x(t_0)$, we can see that $|x(t)| \to 0$ as $t \to \infty$ at a rate, $\lambda$, which is independent of $t_0$. So the condition is sufficient for *global* UAS.

Conversely, assume that the origin is UAS, then there must exist a class $\mathcal{KL}$ function, $\beta : [0, \infty) \times [0, \infty) \to [0, \infty)$, such that

$$|x(t)| \leq \beta(|x(t_0)|, t - t_0), \quad \text{for all } t \geq t_0 \text{ and all } x(t_0) \in \mathbb{R}^n$$

A class $\mathcal{KL}$ function, $\beta(r, s)$, is a continuous function that is continuous and increasing in $r$ with $\beta(0, s) = 0$ and that is also asymptotically decreasing to zero in $s$. It is what we sometimes refer to as a *comparison function*.

Now note that the induced matrix norm of $\mathbf{\Phi}$ has the property

$$\|\mathbf{\Phi}(t; t_0)\| = \max_{|x|=1} |\mathbf{\Phi}(t; t_0)x| \leq \max_{|x|=1} \beta(|x|, t - t_0) = \beta(1, t - t_0)$$

Since $\beta(1, s) \to 0$ as $s \to \infty$, there exists $T > 0$ such that $\beta(1, T) \leq \frac{1}{e}$. For any $t \geq t_0$, let $N$ be the smallest positive integer such that $t \leq t_0 + NT$. Divide the interval $[t_0, t_0 + (N - 1)T]$ into $(N - 1)$ equal subintervals of width $T$. Using the transition matrix' group property we can write

$$\mathbf{\Phi}(t; t_0) = \mathbf{\Phi}(t, t_0 + (N-1)T)\ \mathbf{\Phi}(t_0 + (N-1)T, t_0 + (N-2)T)\ \cdots\ \mathbf{\Phi}(t_0 + T, t_0)$$

and so

$$\begin{aligned} \|\mathbf{\Phi}(t; t_0)\| &\leq \|\mathbf{\Phi}(t, t_0 + (N-1)T)\| \prod_{k=1}^{n-1} \|\mathbf{\Phi}(t_0 + kT, t_0 + (k-1)T)\| \\ &\leq \beta(1, 0) \prod_{k=1}^{N-1} \frac{1}{e} = e\beta(1, 0)e^{-N} \\ &\leq e\beta(1, 0)e^{-(t-t_0)/T} = ke^{-\lambda(t-t_0)} \end{aligned}$$

where $k = e\beta(1, 0)$ and $\lambda = 1/T$. $\diamondsuit$

**Example:** Consider the LTV system

$$\dot{x}(T) = \begin{bmatrix} -1 & \alpha(t) \\ -\alpha(t) & -1 \end{bmatrix} x$$

where $\alpha(t)$ is continuous for all $t \geq 0$. Is the origin uniformly asymptotically stable?

We can solve this by checking a somewhat easier condition suggested by the preceding theorem. In particular, let $V(x) = \frac{1}{2}(x_1^2 + x_2^2)$ and note that

$$\dot{V}(x) = x_1(-x_1 + \alpha(t)x_2) + x_2(-\alpha(t)x_1 - x_2) = -x_1^2 - x_2^2 = -2V$$

In other words, we know that $V(x(t))$ will satisfy the differential equation

$$\dot{V}(t) = -2V(t)$$

which has the solution $V(t) = V(0)e^{-2t}$ for $t \geq 0$. Since $V(x) = \frac{1}{2}|x|^2$ we know this implies the system is uniformly exponentially stable and so by the preceding theorem it must also be UAS.

**Example:** Consider the linear homogeneous system

$$\dot{x}(t) = \begin{bmatrix} -2 & t \\ 0 & -2 \end{bmatrix} x(t)$$

Determine if the origin is uniformly asymptotically stable. We can use our earlier methods to show that this system's state transition matrix is

$$\mathbf{\Phi}(t;\tau) = \begin{bmatrix} e^{-2(t-\tau)} & \frac{t^2-\tau^2}{2}e^{-2(t-\tau)} \\ 0 & e^{-2(t-\tau)} \end{bmatrix} = \begin{bmatrix} 1 & \frac{t^2-\tau^2}{2} \\ 0 & 1 \end{bmatrix} e^{-2(t-\tau)}$$

The matrix norm is clearly dominated by the $e^{-2(t-\tau)}$ term and so we know this system is UAS according to the preceding theorem.

This theorem means that for linear systems, UAS is equivalent to uniform exponential stability (i.e. asymptotic stability where $|x(t)| < ke^{-\lambda t}$). This condition is not as useful as the eigenvalue condition we had for LTI systems because it needs knowledge of the transition matrix that can only be obtained by solving the state equations. In other words, the preceding theorem is of limited value as a "test" for UAS.

We can establish a *converse theorem* for LTV systems. This is done by making considering

$$\mathbf{P}(t) = \int_t^\infty \mathbf{\Phi}^T(\tau, t)\mathbf{Q}(\tau)\mathbf{\Phi}(\tau, t)d\tau$$

and let $\phi(\tau; t, x) = \mathbf{\Phi}(\tau, t)x$ denote the state trajectory with initial condition at time $t$ being $x$. With this notational convention, we can write

$$x^T\mathbf{P}(t)x = \int_t^\infty \phi^T(\tau; t, x)\mathbf{Q}(\tau)\phi(\tau; t, x)d\tau$$

So by the preceding theorem we know that if the equilibrium is uniformly exponentially stable (UAS), then there exist $k > 0$ and $\lambda > 0$ such that

$$\|\mathbf{\Phi}(\tau, t)\| \leq ke^{-\lambda(\tau - t)}$$

Since we assumed $\mathbf{Q}(t)$ is bounded and positive definite, there are constants $c_3$ and $c_4$ such that $c_3\mathbf{I} \leq \mathbf{Q}(t) \leq c_4\mathbf{I}$ and we can then say

$$\begin{aligned} x^T\mathbf{P}(t)x &= \int_t^\infty \phi^T(\tau; t, x)\mathbf{Q}(\tau)\phi(\tau; t, x)d\tau \\ &= \int_t^\infty x^T\mathbf{\Phi}^T(\tau, t)\mathbf{Q}(\tau)\mathbf{\Phi}(\tau, t)xd\tau \\ &\leq \int_t^\infty c_4 \|\mathbf{\Phi}(\tau, t)\|^2 |x|^2 d\tau \\ &\leq \int_t^\infty k^2 e^{-2\lambda(\tau - t)}d\tau c_4|x|^2 \\ &= \frac{k^2c_4}{2\lambda}|x|^2 \end{aligned}$$

So we take $\overline{W}(x) = \dfrac{k^2c_4}{2\lambda}|x|^2$.

On the other hand since there exists $L > 0$ such that $\|\mathbf{A}(t)\| \leq L$ for all time, we can bound the solution from below by

$$|\phi(\tau; t, x)|^2 \geq |x|^2 e^{-2L(\tau - t)}$$

and so

$$\begin{aligned} x^T\mathbf{P}(t)x &\geq \int_t^\infty c_3|\phi(\tau;t,x)|^2 d\tau \\ &\geq \int_t^\infty e^{-2L(\tau-t)}d\tau c_3|x|^2 \\ &= \frac{c_3}{2L}|x|^2 \end{aligned}$$

So we take $\underline{W}(x) = \frac{c_3}{2L}|x|^2$ and what we've established is that

$$\underline{W}(x) = \frac{c_3}{2L}|x|^2 \leq x^T\mathbf{P}(t)x \leq \frac{k^2c_4}{2\lambda}|x|^2 = \overline{W}(x)$$

which establishes the first condition we need for a Lyapunov function.

We now show that the derivative property holds. We use the fact that

$$\frac{\partial}{\partial t}\mathbf{\Phi}(\tau,t) = -\mathbf{\Phi}(\tau,t)\mathbf{A}(t)$$

to show that

$$\begin{aligned} \dot{\mathbf{P}}(t) &= \int_t^\infty \mathbf{\Phi}^T(\tau,t)\mathbf{Q}(\tau)\frac{\partial}{\partial t}\mathbf{\Phi}(\tau,t)d\tau \\ &\quad + \int_t^\infty \left[\frac{\partial}{\partial t}\mathbf{\Phi}^T(\tau,t)\right]\mathbf{Q}(\tau)\mathbf{\Phi}(\tau,t)d\tau - \mathbf{Q}(t) \\ &= -\int_t^\infty \mathbf{\Phi}^T(\tau,t)\mathbf{Q}(\tau)\mathbf{\Phi}(\tau,t)\mathbf{A}(t) \\ &\quad -\mathbf{A}^T(t)\int_t^\infty \mathbf{\Phi}^T(\tau,t)\mathbf{Q}(\tau)\mathbf{\Phi}(\tau,t)d\tau - \mathbf{Q}(t) \\ &= -\mathbf{P}(t)\mathbf{A}(t) - \mathbf{A}^T(t)\mathbf{P}(t) - \mathbf{Q}(t) \end{aligned}$$

which establishes that $V(t,x) = x^T\mathbf{P}(t)x$ is a Lyapunov function. What we have just done is show that if the origin of the LTV system is asymptotically stable, then there must be a Lyapunov function of the form given above. This converse theorem is formally stated in the following theorem.

THEOREM 14. *Let $x = 0$ be the uniformly exponentially stable equilibrium of $\dot{x}(t) = \mathbf{A}(t)x(t)$ where $\mathbf{A}(t)$ is continuous and bounded. Let $\mathbf{Q}(t)$ be continuous bounded symmetric positive definite matrix function of time.*

*Then there is a continuously differentiable bounded positive definite symmetric matrix function, $\mathbf{P}(t)$, that satisfies the matrix differential equation*

$$-\dot{\mathbf{P}}(t) = \mathbf{P}(t)\mathbf{A}(t) + \mathbf{A}^T(t)\mathbf{P}(t) + \mathbf{Q}(t)$$

*and so $V(t,x) = x^T\mathbf{P}(t)x$ is a Lyapunov function for this system.*

## 6. $\mathcal{L}_p$ Stability:

Lyapunov stability is a property of the equilibrium for a system that is not being forced by an unknown exogenous input. For systems with exogenous inputs, the state equation takes the form,

$$\dot{x}(t) = f(x(t), w(t))$$

where $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is now a function of the state $x : \mathbb{R} \to \mathbb{R}^n$ and an applied input signal $w : \mathbb{R} \to \mathbb{R}^m$. One can think of $w$ as a *disturbance*. The disturbance is a signal that we don't know and that fluctuates about a bias or trend line. If $0 = f(0, w(t))$ for any $w$, then the origin is an equilibrium point for the forced system and one can examine the Lyapunov stability of that equilibrium. In general, however, the disturbance is *non-vanishing* in the sense that $f(0, w) \neq 0$ and this means that the origin will not be an equilibrium point and so the Lyapunov stability concept cannot be used.

For systems with non-vanishing perturbations, one often uses a *stability concept* that focuses on the input/output behavior of the system. In particular, we say a forced system is *input/output stable* if all bounded inputs to the system result in a bounded response. "Bounded", in this case, means that signals have finite norms and so the system is viewed as a linear transformation between two normed linear signal spaces. A stable system is then a linear transformation, $\mathbf{G} : \mathcal{L}_{\text{in}} \to \mathcal{L}_{\text{out}}$, that takes any signal, $w \in \mathcal{L}_{\text{in}}$, such that $\|w\|_{\mathcal{L}_{\text{in}}} < \infty$ onto an output signal $\mathbf{G}[w] \in \mathcal{L}_{\text{out}}$ such that $\|\mathbf{G}[w]\|_{\mathcal{L}_{\text{out}}}$ is also finite. It is customary to consider signals that are linear transformations between two $\mathcal{L}_p$ spaces because these spaces are Banach spaces (complete normed linear spaces). Such systems are said to be $\mathcal{L}_p$ stable. The purpose

of this section is to formalize the $\mathcal{L}_p$ stability concept and discuss the special case when $p = 2$.

$\mathcal{L}_p$-stability is defined for systems, $\mathbf{G} : \mathcal{L}_{\text{pe}} \to \mathcal{L}_{\text{pe}}$, that are linear transformations between two *extended* $\mathcal{L}_p$ spaces. In particular, $\mathcal{L}_{\text{pe}}$ is the space of all functions, $w$, such that the truncation of $w$ for any finite $T$

$$w_T(t) = \begin{cases} w(t) & \text{for } t \leq T \\ 0 & \text{otherwise} \end{cases}$$

is in $\mathcal{L}_p$. We say this space is "extended" because it contains all signals in $\mathcal{L}_p$ as well as unbounded signals whose truncations are bounded. For such systems we say $\mathbf{G}$ is $\mathcal{L}_p$ *stable* if and only if there exists a class $\mathcal{K}$ function[5], $\alpha : [0, \infty) \to [0, \infty)$, and a non-negative constant, $\beta$, such that

$$\|\mathbf{G}[w]_T\|_{\mathcal{L}_p} \leq \alpha(\|w_T\|_{\mathcal{L}_p}) + \beta$$

for all $w \in \mathcal{L}_{\text{pe}}$ and $T \geq 0$. The constant $\beta$ is called a *bias*.

Note that in characterizing a system as a linear transformation, we also need to constrain the system to be *causal*. Causality means that the current output is only a function of the past inputs and outputs, a consequence of the forward motion of time. Formally, we define causality using a truncation of the input and output signals. In particular, if we let $w \in \mathcal{L}_p$, then the truncation of $w$ with respect to time instant $T$ is a function $w_T : \mathbb{R} \to \mathbb{R}$ such that

$$w_T(t) = \begin{cases} w(t) & \text{if } t \leq T \\ 0 & \text{otherwise} \end{cases}$$

The dynamical systm $\mathbf{G} : \mathcal{L}_p \to \mathcal{L}_p$ is causal if and only if for any $T \in \mathbb{R}$, we have

$$\mathbf{G}[w_T](t) = \mathbf{G}[w](t), \quad \text{or all } t \leq T$$

Informally, this means that the output of the system prior to time $T$ under the non-truncating input signal $w$ is identical to the system's output prior

[5]A function $\alpha : [0, a) \to [0, \infty)$ is class $\mathcal{K}$ if and only if it is a continuous and increasing with $\alpha(0) = 0$.

to time $T$ under the truncated input $w_T$. Since $w_T$ is zero for $t > T$, this means that nonzero inputs after time $T$ have no impact on outputs prior to $T$. In other words, the future inputs have no impact on the past outputs.

Lyapunov analysis is often just concerned with declaring whether or not the equilibrium is stable. But for $\mathcal{L}_p$ stability, one talks about how well the "disturbance", $w$, is attenuated at the system's output and this degree of attenuation is characterized through the concept of the system's *gain*. In particular, we say that $\mathbf{G} : \mathcal{L}_{\text{pe}} \to \mathcal{L}_{\text{pe}}$ is *finite-gain $\mathcal{L}_p$ stable* if there exist $\gamma > 0$ such that

$$\|\mathbf{G}[w]_T\|_{\mathcal{L}_p} \leq \gamma \|w_T\|_{\mathcal{L}_p} + \beta$$

This constant $\gamma$ is called a *gain*.

Note that if there is any other $\gamma_1 > \gamma$, then $\gamma_1$ is also a gain. We want a notion of gain as a property of the *system* that is some sense uniquely defined. This is done by taking the infimum over all inputs, $w$, of those scalars that can be gains. We call this the $\mathcal{L}_p$ *induced gain* of the system that is formally defined as

$$\|\mathbf{G}\|_{\mathcal{L}_p-\text{ind}} := \inf \left\{ \gamma \,:\, \|(\mathbf{G}[w])_T\|_{\mathcal{L}_p} \leq \gamma \|w_T\|_{\mathcal{L}_p} + \beta, \quad \text{for all } w \in \mathcal{L}_p \text{ and } T \geq 0 \right\}$$

Note that if the bias, $\beta$, is zero then the above gain is identical to the $\mathcal{L}_2$ and $\mathcal{L}_\infty$ defined we discussed in earlier chapters.

The prior lectures provided an explicit formula for the $\mathcal{L}_\infty$ induced gain of an LTI system with a known impulse response function. But the $\mathcal{L}_2$-induced gain we derived for LTI systems required finding the maximum of the system's gain-magnitude function, $|\mathbf{G}(j\omega)|$. Finding this maximum can be very difficult to do if the system has many sharp resonant peaks; which is usually the case for mechanical systems with a large number of vibrational modes. So the rest of this section presents an alternative way

of determining the $\mathcal{L}_2$-induced gain that is computationally efficient. The following theorem will play an important role in this approach.

THEOREM 15. *Suppose* $\mathbf{G} \stackrel{s}{=} \left[\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & 0 \end{array}\right]$ *is a system realization with* $\mathbf{A}$ *being Hurwitz, then* $\|\mathbf{G}\|_{\mathcal{L}_2-\mathrm{ind}} < \gamma$ *if and only if the matrix*

$$\mathbf{H} = \begin{bmatrix} \mathbf{A} & \frac{1}{\gamma^2}\mathbf{B}\mathbf{B}^T \\ -\mathbf{C}^T\mathbf{C} & -\mathbf{A}^T \end{bmatrix}$$

*has no eigenvalues on the* $j\omega$*-axis.*

**Proof:** Let $\mathbf{\Phi}(s) = \gamma^2\mathbf{I} - \mathbf{G}^*(s)\mathbf{G}(s)$. It should be apparent that because $\|\mathbf{G}\|_{\mathcal{L}_2-\mathrm{ind}} = \max_\omega |\mathbf{G}(j\omega)|$ then $\|\mathbf{G}\|_{\mathcal{L}_2-\mathrm{ind}} < \gamma$ if and only if $\mathbf{\Phi}(j\omega) > 0$ for all $\omega \in \mathbb{R}$. This means that $\mathbf{\Phi}(s)$ has no zeros on the imaginary axis, or rather than $\mathbf{\Phi}^{-1}(s)$ has no poles on the imaginary axis.

One can readily verify that a state space realization for $\mathbf{\Phi}^{-1}(s)$ is

$$\mathbf{\Phi}^{-1}(s) \stackrel{s}{=} \left[\begin{array}{c|c} \mathbf{H} & \begin{bmatrix} \frac{1}{\gamma^2}\mathbf{B} \\ \mathbf{0} \end{bmatrix} \\ \hline \begin{bmatrix} \mathbf{0} & \frac{1}{\gamma^2}\mathbf{B}^T \end{bmatrix} & \frac{1}{\gamma^2} \end{array}\right]$$

If $\mathbf{H}$ has no imaginary eigenvalues, then clearly $\mathbf{\Phi}^{-1}(s)$ has no imaginary poles.

So let $j\omega_0$ be an eigenvalue of $\mathbf{H}$. This means there is $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \neq 0$ such that

$$\begin{aligned} 0 &= (j\omega_0\mathbf{I} - \mathbf{H})x \\ &= \begin{bmatrix} j\omega_0\mathbf{I} - \mathbf{A} & -\frac{1}{\gamma^2}\mathbf{B}\mathbf{B}^T \\ \mathbf{C}^T\mathbf{C} & j\omega_0\mathbf{I} + \mathbf{A}^T \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \end{aligned}$$

which means

$$(j\omega_0\mathbf{I} - \mathbf{A})x_1 = \frac{1}{\gamma^2}\mathbf{B}\mathbf{B}^T x_2 \tag{34}$$

$$(j\omega_0\mathbf{I} + \mathbf{A}^T)x_2 = -\mathbf{C}^T\mathbf{C}x_1 \tag{35}$$

The mode associated with this eigenvalue will be $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} e^{j\omega_0 t}$. For this mode not to appear in the system's output, we would require for all $t$ that

$$0 = \begin{bmatrix} 0 & \frac{1}{\gamma^2}\mathbf{B}^T \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} e^{j\omega_0 t} = \frac{1}{\gamma^2}\mathbf{B}^T x_2 e^{m\omega_0 t}$$

This can only occur in if $\mathbf{B}^T x_2 = 0$, which when we insert this into equations (34-35) gives

$$\begin{aligned} (j\omega_0 \mathbf{I} - \mathbf{A})x_1 &= 0 \\ (j\omega_0 \mathbf{I} + \mathbf{A}^T)x_2 &= -\mathbf{C}^T\mathbf{C}x_1 \end{aligned}$$

The first equation implies $x_1 = 0$ since $\mathbf{A}$ is Hurwitz, and inserting this into the second equation implies $x_2 = 0$. So we've shown that if the $j\omega_0$-mode is activated and it does not appear at the output, then $x = 0$, which cannot happen. So in this case $j\omega_0$ cannot be an eigenvalue of $\mathbf{H}$. A similar case can be made if we require that the applied input not excite those $x_2$ components in the $j\omega_0$-mode. $\diamondsuit$

The preceding theorem is useful because it uses an eigenvalue test to certify whether $\|\mathbf{G}\|_{\mathcal{L}_2-\text{ind}} < \gamma$. We refer to this test as an algorithmic *oracle* since for a given state space realization it can easily declare whether or not a given $\gamma$ is an upper bound on the induced gain. In particular, we use this oracle as the basis of a binary or bisection search to efficiently search for the minimum $\gamma$. This bisection algorithm is stated below.

(1) Select an upper and lower bound, $\gamma_u$ and $\gamma_\ell$, respectively, such that

$$\gamma_\ell \leq \|\mathbf{G}\|_{\mathcal{L}_2-\text{ind}} \leq \gamma_u$$

(2) If $\dfrac{\gamma_u - \gamma_\ell}{\gamma_\ell} < \epsilon$ where $\epsilon$ is a specified error tolerance, then STOP and declare $\dfrac{\gamma_u + \gamma_\ell}{2}$ as the induced gain. Otherwise continue

(3) Set $\gamma = \dfrac{\gamma_u + \gamma_\ell}{2}$

(4) Form the Hamiltonian matrix $\mathbf{H}$ and compute its eigenvalues

- if no eigenvalues are purely imaginary, then set $\gamma_\ell = \gamma$ and go to step 2.
- If any eigenvalue is imaginary then set $\gamma_u = \gamma$ and go to step 2.

This algorithm works in a very simple manner. It assumes we already know that the induced gain is bounded between the two initial guesses, $\gamma_\ell$ and $\gamma_u$. The value of $\gamma$ that we check is chosen halfway between $\gamma_\ell$ and $\gamma_u$, thereby dividing the region of uncertainty in half. Let $\gamma_0$ denote this initial guess. The eigenvalue test tells us whether the initial guess is or is not an upper bound on the induced gain. If it is an upper bound, then we know $\gamma_\ell < \|\mathbf{G}\|_{\mathcal{L}_2-\text{ind}} < \gamma_0$ and we can reset the upper bound to $\gamma_u = \gamma_0$. If the eigenvalue test tells us $\gamma_0$ is not an upper bound on the induced gain, then we know $\gamma_0 < \|\mathbf{G}\|_{\mathcal{L}_2-\text{ind}} < \gamma_u$ and we can reset the lower bound to $\gamma_\ell = \gamma_0$. With this reset, the interval of uncertainty around the induced gain, $[\gamma_\ell, \gamma_u]$ is cut in half. We then repeat this game with the smaller uncertainty interval. What this algorithm guarantees is that we will determine the induced gain to an accuracy of $\dfrac{\gamma_u - \gamma_\ell}{2^n}$ after $n$ recursions. So for a specified tolerance level, $\epsilon$, we can actually determine how many recursions are needed to complete the search.

**Example:** Consider the LTI system

$$\mathbf{G} \stackrel{s}{=} \left[\begin{array}{cc|c} 0 & 1 & 0 \\ -\omega_n^2 & -0.1 & 1 \\ \hline 1 & 0 & 0 \end{array}\right]$$

Determine the system's induced $\mathcal{L}_2$ gain.

We know the transfer function for this system is $\mathbf{G}(s) = \frac{1}{s^2+0.1s+\omega_n^2}$ and so

$$|\mathbf{G}(j\omega)|^2 = \frac{1}{(\omega_n^2 - \omega^2)^2 + 0.01\omega^2}$$

Computing the first derivative and setting it equal to zero yields,

$$2(\omega_n^2 - \omega) = 0.01 = 0$$

The solution, $\omega_0$, is the peak in the gain magnitude function and satisfies the quadratic equation

$$\omega_0^2 + \omega_n^2 - \frac{0.01}{2}$$

which has a positive solution positive for $\omega_n^2 > 0.01/2$. For $\omega_n^2 \leq 0.01/2$, the function is monotone decreasing which means the peak occurs for $\omega_0 = 0$. We therefore see that

$$\|\mathbf{G}\|_{\mathcal{L}_2-\text{ind}} = |\mathbf{G}(j\omega_0)|^2 = \begin{cases} \frac{1}{(0.01/2)^2+(0.01)(0.01/2+\omega_n^2)} \approx \frac{100}{\omega_n^2} & \text{for } \omega_n > 0.01/2 \\ \frac{1}{\omega_n^2} & \text{for } \omega_n^2 \leq 0.01/2 \end{cases}$$

**Example:** Consider a system with the following state space realization

$$\mathbf{G} \stackrel{s}{=} \left[\begin{array}{cccccccc|c} -0.016 & 16.19 & 0 & 0 & 0 & 0 & 0 & 0 & 1.30 \\ -16.19 & -0.162 & 0 & 0 & 0 & 0 & 0 & 0 & 0.01 \\ 0 & 0 & -0.01058 & 10.58 & 0 & 0 & 0 & 0 & 1.1 \\ 0 & 0 & -10.58 & -0.0106 & 0 & 0 & 0 & 0 & -0.09 \\ 0 & 0 & 0 & 0 & -0.004 & 3.94 & 0 & 0 & -0.22 \\ 0 & 0 & 0 & 0 & -3.94 & -0.004 & 0 & 0 & -0.0019 \\ 0 & 0 & 0 & 0 & 0 & 0 & -0.00056 & 0.568 & 0.074 \\ 0 & 0 & 0 & 0 & 0 & 0 & -0.568 & -0.0006 & 10^{-4} \\ \hline 2 \times 10^{-5} & -0.0025 & -0.0075 & -0.095 & -0.00032 & 0.0366 & 0.00016 & -0.127 & 0 \end{array}\right]$$

Plot the system's frequency response function at $100$ equally spaced points between $0.1$ to $100$ rad/sec. Use that plot to estimate the $\mathcal{L}_2$ induced gain of the system. Then use the preceding bisection algorithm to estimate the induced gain.

The frequency response plotted using MATLAB is shown in Fig. 1. This plots shows points at 100 sample points between $0.1$ and $100$ rad/sec (blue asterisks) as well as a more finely sampled plot with $10,000$ sample points (shown in solid red line). The peak found by the coarse sampling was $1.0822$ ($0.3671$ dB). So one would estimate the induced gain to be $1.0822$. The peak found by the more finely sampled plot was $4.8912$ ($13.7883$ dB). If we had used the preceding bisection algorithm, we would have found the actual peak to be $8.2664$ ($18.3463$ dB). The induced gain estimated from the gain-magnitude plot depended greatly on how finely we sampled the plot. The problem here was that we have no way of relating the sampling interval

to the error in our estimate of the induced gain. The bisection algorithm, on the other hand, does provide a relationship between the number of recursions and the accuracy of the estimate. In particular, one can guarantee that with an initial guess between $0$ and $10$ one would only need $10$ recursions to get within $0.01$ of the actual gain.

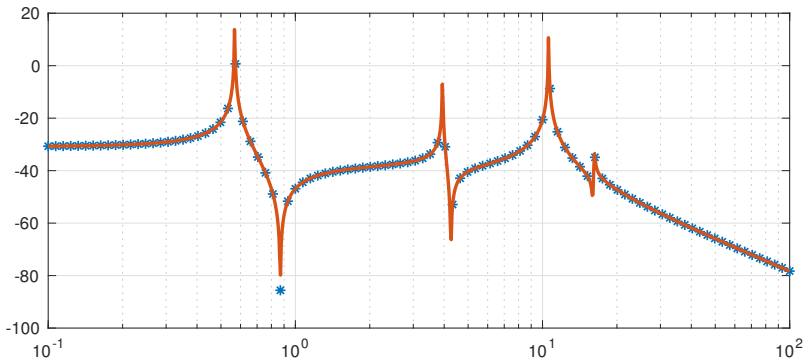



FIGURE 1. Sampled Gain magnitude being used to estimate L2 induced gain

Since Lyapunov stability is such an important concept, it is useful to characterize the relationship between $\mathcal{L}_p$ stability and Lyapunov stability. This is done in the following theorem

THEOREM 16. *Consider the input-output system* $\dot{x}(t) = f(x(t), w(t))$ *and* $y(t) = h(x(t), w(t))$ *where the origin is an exponentially stable equilibrium of* $\dot{x}(t) = f(x(t), 0)$. *Assume there exist positive constants* $L$, $r$, $r_w$, $\eta_1$, *and* $\eta_2$ *such that*

$$\begin{aligned} |f(x,w) - f(x,0)| &\leq L|w| \\ |h(x,w)| &\leq \eta_1|x| + \eta_2|w| \end{aligned}$$

*for all $|x| < r$ and $|w| < r_w$. If there exists a $C^1$ function $V : \mathbb{R}^n \to \mathbb{R}$ and non-negative constants $c_1$, $c_2$, $c_3$, and $c_4$ such that*

$$\begin{aligned} c_1|x|^2 &\leq V(x) \leq c_2|x|^2 \\ \dot{V}(x,0) &\leq -c_3|x|^2 \\ \left|\frac{\partial V}{\partial x}\right| &\leq c_4|x| \end{aligned}$$

*then the system is finite gain $\mathcal{L}_p$-stable.*

**Proof:** Consider $\dot{V}$ along trajectories of the forced system

$$\dot{V} = \frac{\partial V}{\partial x} f(x,0) + \frac{\partial V}{\partial x}[f(x,w) - f(x,0)]$$

with the given bounds and the Lipschitz constant, $L$, for $f$ we get

$$\dot{V} \leq -c_3|x|^2 + c_4 L|x||w|$$

Take $W(t) = \sqrt{V(x(t))}$ and note that

$$\dot{W} = \frac{\dot{V}}{2\sqrt{V}}$$

We obtain the following differential inequality

$$\dot{W} \leq -\frac{1}{2}\frac{c_3}{c_2} W + \frac{c_4 L}{2\sqrt{c_1}}|w(t)|$$

By the *comparison principle* we can therefore conclude that

$$W(t) \leq e^{-\frac{c_3}{2c_2}t} W(0) + \frac{c_4 L}{2\sqrt{c_1}} \int_0^t e^{-(t-s)\frac{c_3}{2c_2}} |w(s)| ds$$

which implies that

$$|x(t)| \leq \frac{c_2}{c_1}|x_0| e^{-\frac{c_3}{2c_2}t} + \frac{c_4 L}{2c_1} \int_0^t e^{-(t-s)\frac{c_3}{2c_2}} |w(s)| ds$$

and so we can conclude that if $|x_0| \leq \frac{r}{2}\sqrt{\frac{c_1}{c_2}}$ and $\|w\|_{\mathcal{L}_\infty} \leq \frac{c_1 c_3 r}{2c_2 c_4 L}$, then $|x(t)| \leq r$ for all time.

This means that the bound on $h$ holds for all time and so

$$|y(t)| \leq k_1 e^{-at} + k_2 \int_0^t e^{-a(t-s)} |w(s)| ds + k_3 |w(t)|$$

where $k_1 = \sqrt{\frac{c_1}{c_2}}|x_0|\eta_1$, $k_2 = \frac{c_4 L \eta_1}{2c_1}$, $k_3 = \eta_2$, and $a = \frac{c_3}{2c_2}$. Assign to each of these terms in the above equation a signal $y_1$, $y_2$, and $y_3$. If $w \in \mathcal{L}_{\rm pe}$ with $\|w\|_{\mathcal{L}_\infty}$ sufficiently small then for any $T > 0$

$$\|y_{2T}\|_{\mathcal{L}_p} \leq \frac{k_2}{a}\|w_T\|_{\mathcal{L}_p}, \quad \|y_{3T}\|_{\mathcal{L}_p} \leq k_3\|w_T\|_{\mathcal{L}_p}, \quad \|y_{1T}\|_{\mathcal{L}_p} \leq k_1\rho$$

where $\rho = \begin{cases} 1 & \text{if } p = \infty \\ (1/ap)^{1/p} & \text{otherwise} \end{cases}$. So that since

$$\|y_T\|_{\mathcal{L}_p} \leq \|y_{1T}\|_{\mathcal{L}_p} + \|y_{2T}\|_{\mathcal{L}_p} + \|y_{3T}\|_{\mathcal{L}_p}$$

we can use the above bounds to conclude

$$\begin{aligned} \|y_T\|_{\mathcal{L}_p} &\leq k_1\rho + \frac{k_2}{a}\|w_T\|_{\mathcal{L}_p} + k_3\|w_T\|_{\mathcal{L}_p} \\ &= \left(\frac{k_2}{a} + k_3\right)\|w_T\|_{\mathcal{L}_p} + k_1\rho = \gamma\|w_T\|_{\mathcal{L}_p} + \beta \end{aligned}$$

which identify the finite gain and bias for this system. $\diamondsuit$

For linear time-invariant systems, since the existence of a Lyapunov function is necessary and sufficient for asymptotic stability of the state, then it is much easier to see that the system will be $\mathcal{L}_p$ stable. The converse, however is not true. Let us consider,

$$\begin{aligned} \dot{x}(t) &= \mathbf{A}x(t) + \mathbf{B}w(t) \\ y(t) &= \mathbf{C}x(t) \end{aligned}$$

If the origin is asymptotically stable, then it is exponentially stable and so there exist $K > 0$ and $\gamma > 0$ such that

$$e^{\mathbf{A}t}x_0 \leq Ke^{-\gamma t}$$

for $t \geq 0$. Since the impulse response is $\mathbf{C}e^{\mathbf{A}t}\mathbf{B}u(t)$, it is easy to see that this must be integrable and so if the state-based system is exponentially stable, it must also be $\mathcal{L}_p$ stable.

The converse may not be true as is seen in the following example. Let us consider the state-based system,

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} w(t)$$

$$y(t) = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}$$

This system is clearly unstable since the $\mathbf{A}$ matrix has one eigenvalue with a positive real part. This system however is $\mathcal{L}_p$ stable since its transfer function is $\frac{1}{s+1}$ which only has a single pole on the left hand side of the complex plane. The reason why this system is $\mathcal{L}_p$ stable but not asymptotically stable is because the unstable mode of the state-based system is not observed at the system's output. Moreover, it is not influenced by the input either. State-based models that have this property are said to be *uncontrollable* and *unobservable*.

CHAPTER 4

# Controllability and Observability

The stabilization problem is concerned with keeping the system state in a neighborhood of an equilibrium. A companion problem is how one steers the state to that neighborhood in the first place. In particular for a continuous-time linear system

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

or a discrete-time system

$$x(k+1) = \mathbf{A}(k)x(k) + \mathbf{B}(k)u(k), \quad x(k_0) = x_0$$

how do we select the controlled input $u$ so the state trajectory starting at time $t_0$ ($k_0$) in state $x_0$ reaches a given operating point in finite time. This is called the *reachability* problem. The finite time nature of the problem's statement makes it distinct from the stabilization problem we considered earlier. We will find it useful to consider two versions of this problem.

- The *reachability* problem seeks a control input that drives the state to a desired point from the origin in finite time.
- The *controllability* problem seeks a control that steers the system state to the origin for any initial state in finite time.

For continuous-time systems, these two versions of the concept are equivalent. For discrete-time systems, these concepts are not equivalent since time cannot flow freely in both directions for a discrete-time system.

Dual to the notion of reachability/controllability is *observability*. Observability asks whether one can determine the system's state $x$ at time $t_0$ if we have access to a finite interval of inputs $u_{[t_0,t_0+T]}$ and output $y_{[t_0,t_0+T]}$. A

related concept known as *constructibility* asks whether the state $x$ at time $t_0$ can be predicted based on inputs for a finite time interval, $[t_0 - T, t_0]$, prior to $t_0$. This chapter discusses all four of these finite-time concepts.

## 1. Controllability/Reachability Definitions

Let us consider an inhomogeneous time-varying linear system whose state equations are

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

Informally, we want to know if there is an input, $u$, that transfers the state from $x_0$ at time $t_0$ to a specified state, $x_1$, by some *finite* time $t_1 > t_0$. If such an input exists then we must have

$$x_1 = \mathbf{\Phi}(t_1; t_0)x_0 + \int_{t_0}^{t_1} \mathbf{\Phi}(t_1; \tau)\mathbf{B}(\tau)u(\tau)d\tau$$

It will be convenient to rewrite this equation as

$$\widehat{x}_1 \stackrel{\text{def}}{=} x_1 - \mathbf{\Phi}(t_1; t_0)x_0 = \int_{t_0}^{t_1} \mathbf{\Phi}(t_1; \tau)\mathbf{B}(\tau)u(\tau)d\tau$$

Rewriting the equation this way suggests that the problem of transfering the state from $(x_0, t_0)$ to $(x_1, t_1)$ is equivalent to transfering the state from the origin, $(0, t_0)$, to $(\widehat{x}_1, t_0)$. For this reason, the formal definition of reachability is defined with respect to steering the state from the origin to a specified target state, $x_1$. So we say that state $x_1$ is reachable at $t_1$ if and only if there exists a finite $t_0 < t_1$ and input $u : [t_0, t_1] \to \mathbb{R}^m$ such that

$$x_1 = \int_{t_0}^{t_1} \mathbf{\Phi}(t_1; \tau)\mathbf{B}(\tau)u(\tau)d\tau \tag{36}$$

The set of all reachable states for a given system pair, $(\mathbf{A}(t), \mathbf{B}(t))$ forms a linear space. The *reachable at $t_1$ subspace* of $(\mathbf{A}(t), \mathbf{B}(t))$ is

$$\mathcal{R}_r^{t_1} \stackrel{\text{def}}{=} \left\{ x_1 \in \mathbb{R}^n \;:\; \begin{array}{c} \text{for some finite } t_0 < t_1 \text{ and there is an input } u \\ \text{under which } x_1 = \displaystyle\int_{t_0}^{t_1} \mathbf{\Phi}(t_1; \tau)\mathbf{B}(\tau)u(\tau)d\tau \end{array} \right\}$$

The *system* is reachable at $t_1$ if $\mathcal{R}_r^{t_1} = \mathbb{R}^n$. If the union of all reachable subspaces at $t_1$, i.e. $\bigcup_{t_1} \mathcal{R}_r^{t_1}$, consists of the entire state space, then the system is reachable. In plain language, this means a system is reachable if every state can be reached in finite time from the origin. If the system is state reachable, then it is customary to say that $(\mathbf{A}(t), \mathbf{B}(t))$ forms a *reachable pair*.

The *controllability* concept reverses time to identify those states that can be driven to the origin in finite time by an input. Usually the origin is the system's equilibrium point and so we say a state $x_0$ is *controllable* at time $t_0$ if there exists a finite time $t_1 > t_0$ and input $u : [t_0, t_1] \to \mathbb{R}^m$ that transfers the state from $x_0$ at $t_0$ to the origin at $t_1$. In other words, for some finite time $t_1 > t_0$ there will be an input such that the state at time $t_1$ is at the origin, i.e. $0 = x(t_1\,;\, x_0, t_0)$. This is equivalent to saying $x_0$ is controllable at $t_0$ if there exists a finite time $t_1 > t_0$ and an input $u$ such that

$$-\mathbf{\Phi}(t_1; t_0)x_0 = \int_{t_0}^{t_1} \mathbf{\Phi}(t_1; \tau)\mathbf{B}(\tau)u(\tau)d\tau$$

Note that because the transition matrix, $\mathbf{\Phi}$, is invertible for continuous-time systems, so the preceding integral equation is equivalent to

$$-x_0 = \int_{t_0}^{t_1} \mathbf{\Phi}(t_0; \tau)\mathbf{B}(\tau)u(\tau)d\tau \tag{37}$$

which is very similar to the integral equation (36) used to characterize reachability.

As we did for reachability, we can define the *controllable at* $t_0$ *subspace* as

$$\mathcal{R}_c^{t_0} \stackrel{\text{def}}{=} \left\{ x_0 \in \mathbb{R}^n \; : \; \begin{array}{l} \text{there exists finite } t_1 > t_0 \text{ and an input } u \\ \text{such that } -x_0 = \displaystyle\int_{t_0}^{t_1} \mathbf{\Phi}(t_0; \tau)\mathbf{B}(\tau)u(\tau)d\tau \end{array} \right\}$$

We say the system is *controllable* at $t_0$ if $\mathcal{R}_c^{t_0} = \mathbb{R}^n$. The *system is controllable* if $\bigcup_{t_0} \mathcal{R}_c^{t_0} = \mathbb{R}^n$. Note that the integral conditions for controllability (37) and reachability (36) are essentially identical for continuous-time

systems. One may therefore conjecture that a continuous-time system is controllable if and only if it is reachable. Note that the non-singular nature of the continuous-time system's transition matrix played a major role in forming this conjecture. For discrete-time systems

$$x(k+1) = \mathbf{A}(k)x(k) + \mathbf{B}(k)u(k)$$

the transition matrix $\mathbf{\Phi}(k;k_0) = \prod_{i=k_0}^{k} \mathbf{A}(i)$ may not be invertible. So in general, we cannot make the same conjecture. In other words, controllability and reachability are not equivalent for discrete-time systems and this fact is why we draw a distinction between the two notions.

The equivalence of controllability and reachability for continuous-time systems may be proven in a "formal" manner with a *functional*, $\mathbf{L}[\cdot;t_0,t_1] : \mathcal{L}_{\text{pe}} \to \mathbb{R}^n$ that takes values

$$\mathbf{L}[u;t_0,t_1] = \int_{t_0}^{t_1} \mathbf{\Phi}(t_1;\tau)\mathbf{B}(\tau)u(\tau)d\tau \tag{38}$$

With this linear functional, we can see that equation (36) asserts $x \in \mathbb{R}^n$ is reachable at $t_1$ if and only if there exists a finite $t_0 < t_1$ such that $x$ lies in the range space of $\mathbf{L}$. In other words

$$(x,t_0) \text{ is reachable at } t_1 \quad \Leftrightarrow \quad x \in \text{Range}(\mathbf{L}[\cdot\,;\,t_0,t_1]) \text{ for some input } u$$

The controllability condition in equation (37) asserts that $x$ is controllable from $t_0$ if and only if there exists finite $t_1 > t_0$ and input $u$ such that

$$\begin{aligned} x &= -\int_{t_0}^{t_1} \mathbf{\Phi}(t_0;\tau)\mathbf{B}(\tau)u(\tau)d\tau \\ &= \int_{t_1}^{t_0} \mathbf{\Phi}(t_0;\tau)\mathbf{B}(\tau)u(\tau)d\tau = \mathbf{L}[u\,;\,t_1,t_0] \end{aligned}$$

which is the same as saying $x$ is controllable from $t_0$ if there exists finite $t_1 > t_0$ such that

$$(x,t_1) \text{ is controllable at } t_0 \quad \Leftrightarrow \quad x \in \text{Range}(\mathbf{L}[\cdot\,;\,t_1,t_0]) \text{ for some input } u$$

The subspaces $\text{Range}(\mathbf{L}[\cdot\,;\,t_1,t_0])$ and $\text{Range}(\mathbf{L}[\cdot\,;\,t_0,t_1])$ are readily shown to be equivalent thereby allowing us to conclude a state $x$ is controllable at $x_0$ if and only if it is also reachable at $t_1$. This argument "formalizes" the observation leading to our original conjecture. The preceding argument with the linear functional $\mathbf{L}$ simply allows us to formally verify that our conjecture is indeed true. This conjecture can now be formalized into the following theorem

THEOREM 17. *Consider the LTV system,* $\dot{x}(t) = \mathbf{A}(t)x(t) + \mathbf{B}(t)u(t)$. *The state* $x \in \mathbb{R}^n$ *is controllable at* $t_0$ *if and only if it is reachable at* $t_1$.

## 2. Conditions for Reachability/Controllability

Equations (36) and (37) provide conditions that can verify whether a state $x$ is reachable/controllable. But these conditions are difficult to verify because we need to determine the range space of a linear functional equation. Determining whether states are reachable/controllable is of particular interest to safety critical systems. Systems can become "unsafe" if their state enters a "forbidden" region of the state space in finite time. We therefore need to identify conditions for controllability and reachability that are easier to verify than the integral equations given in the preceding section.

The *reachability gramian* of the continuous-time system

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

is an $n \times n$ matrix-valued function

$$\mathbf{W}_r(t_0;t_1) = \int_{t_0}^{t_1} \mathbf{\Phi}(t_1;\tau)\mathbf{B}(\tau)\mathbf{B}^T(\tau)\mathbf{\Phi}^T(t_1;\tau)d\tau$$

where $\mathbf{\Phi}(t_1;t_0)$ is the system's transition matrix.

Recall that $(x,t_0)$ is reachable at $t_1$ if and only if $x$ lies in the range space of the linear functional $\mathbf{L}[u\,;\,t_0,t_1]$ for some input $u$. Verifying this condition, however, is difficult because we must look through all inputs $u$, to see if its true for a single one. We will show that the Range space of

the linear functional, $\mathbf{L}$, is equivalent to the range space of the reachability gramian, $\mathbf{W}_r$, thereby providing another way of verifying whether $x$ is reachable without having to find $u$ directly.

Let us first verify $\text{Range}(\mathbf{W}_r) \subset \text{Range}(\mathbf{L})$. Let $x_1 \in$ Range$(\mathbf{W}_r)$. This means there exists $\eta_1 \in \mathbb{R}^n$ such that $\mathbf{W}_r\eta_1 = x_1$. So choose

$$u(\tau) = \mathbf{B}^T(\tau)\mathbf{\Phi}^T(t_1\,;\tau)\eta_1$$

for $\tau \leq t_1$, then

$$\begin{aligned}\mathbf{L}[u;t_0,t_1] &= \left[\int_{t_0}^{t_1} \mathbf{\Phi}(t_1;\tau)\mathbf{B}(\tau)\mathbf{B}^T(\tau)\mathbf{\Phi}^T(t_1;\tau)d\tau\right]\eta_1 \\ &= \mathbf{W}_r(t_0,t_1)\eta_1 = x_1\end{aligned}$$

This means that $x_1 \in \text{Range}(\mathbf{L})$ for the chosen $u$. We can therefore conclude that $\text{Range}(\mathbf{W}_r) \subset \text{Range}(\mathbf{L})$.

We can now prove the other direction, namely $\text{Range}(\mathbf{L}) \subset \text{Range}(\mathbf{W}_r)$. Let $x_0 \in \text{Range}(\mathbf{L})$, so there exists some control input, $u$, such that $\mathbf{L}\,[u\,;\,t_0,t_1] = x_1$. Let us assume, however, that $x_1 \notin \text{Range}(\mathbf{W}_r)$. We are going to use the fundamental theorem of linear algebra to show that this cannot occur. From the fundamental theorem of linear algebra, we know that for any linear transformation (matrix), $\mathbf{A}$, we have

$$\ker(\mathbf{A}) = (\text{Range}(\mathbf{A}^T))^{\perp}$$

Since $\mathbf{W}_r$ is a symmetric matrix, we can conclude

$$\text{Range}(\mathbf{W}_r) = (\ker(\mathbf{W}_r))^{\perp}$$

So for any $w \in \text{Range}(\mathbf{W}_r)$ and $v \in \ker(\mathbf{W}_r)$ we would have $w^Tv = 0$. So let us take our $x_1$ and rewrite it as

$$x_1 = x_1' + x_1''$$

where $x_1' \in \text{Range}(\mathbf{W}_r)$ and $x$" $\in \ker(\mathbf{W}_r)$. Note that $x_1'' \neq 0$ because we assumed $x_1 \notin \text{Range}(\mathbf{W}_r)$. There exists, therefore an $x_2 \in \ker(\mathbf{W}_r)$ such that $x_2^Tx_1'' \neq 0$, which would imply $x_2^Tx_1 \neq 0$.

But if this nonzero $x_2$ is in $\ker(\mathbf{W}_r)$ then we can readily see

$$\begin{aligned} x_2^T \mathbf{W}_r x_2 &= 0 = \int_{t_0}^{t_1} (x_2^T \mathbf{\Phi B})(x_2^T \mathbf{\Phi B})^T d\tau \\ &= \int_{t_0}^{t_1} \left| x_2^T \mathbf{\Phi}(t_1;\tau)\mathbf{B}(\tau) \right|^2 d\tau \end{aligned}$$

which would mean $x_2^T \mathbf{\Phi}(t_1;\tau)\mathbf{B}(\tau) = 0$ for any $\tau \in [t_0, t_1]$. This observation would imply

$$x_2^T x_1 = x_2^T \mathbf{L}[u; t_0, t_1] = \int_{t_0}^{t_1} x_2^T \mathbf{\Phi}(t_1;\tau)\mathbf{B}(\tau)u(\tau)d\tau = 0$$

where $u$ is the input taking the state to $x_1$ from the origin. This last equation, however, contradicts our earlier observation that $x_2^T x_1 \neq 0$. This contradiction arose from our assumption that $x_1 \notin \text{Range}(\mathbf{W}_r)$ and so the contraction implies $x_1 \in \text{Range}(\mathbf{W}_r)$. Since our original choice of $x_1 \in \text{Range}(\mathbf{L})$ was arbitrary, we can conclude $\text{Range}(\mathbf{L}) \subset \text{Range}(\mathbf{W}_r$. The preceding arguments allows us to conclude $\text{Range}(\mathbf{L}) = \text{Range}(\mathbf{W}_r)$.

Since the range spaces of $\mathbf{L}$ and $\mathbf{W}_r$ are the same, it means that we can verify that a state $x_1$ is reachable at $t_1$ if and only if it lies in the the range space of $\mathbf{W}_r(t_0, t_1)$ for some finite $t_0 < t_1$. This condition is much easier to certify than the condition in integral equation (36) because $\mathbf{W}_r$ is only a matrix that is independent of the input $u$. Another important outcome of our preceding discussion is that it identifies one particular input $u$ that steers the state from the origin to $x_1$. In particular that transfer must be achieved by the input $u(t) = \mathbf{B}^T(t)\mathbf{\Phi}^T(t_1, t)\eta_1$ that we used to form the reachability gramian from the linear functional $\mathbf{L}$. These observations can be summarized in the following theorem.

THEOREM 18. *For the LTV system $\dot{x}(t) = \mathbf{A}(t)x(t) + \mathbf{B}(t)u(t)$, then the state $x_1$ is reachable at $t_1$ if and only if there exists a finite $t_0 < t_1$ such that*

$$x_1 \in \text{Range}(\mathbf{W}_r(t_0, t_1))$$

*Moreover, one such input that achieves this transfer is*

$$u(t) = \mathbf{B}^T(t)\mathbf{\Phi}^T(t_1, t)\eta_1$$

*where $\eta_1$ is a solution of the linear algebraic equation*

$$\mathbf{W}_r(t_0, t_1)\eta_1 = x_1$$

In view of the preceding theorem one can assert that all states can be reached at $t_1$ if and only if for some $t_0 < t_1$ we have $\mathrm{Range}(\mathbf{W}_r(t_0, t_1)) = \mathbb{R}^n$. This can only be true if $\mathrm{rank}(\mathbf{W}_r(t_0, t_1)) = n$ are rather that $\det \mathbf{W}_r(t_0, t_1) \neq 0$. These observations are summarized in the following theorem

THEOREM 19. *$(\mathbf{A}(t), \mathbf{B}(t))$ is a reachable pair at $t_1$ if and only if there exists $t_0 < t_1$ such that $\det(\mathbf{W}_r(t_0, t_1)) \neq 0$.*

**Example:** Consider the LTV system

$$\dot{x}(t) = \begin{bmatrix} -1 & e^{2t} \\ 0 & -1 \end{bmatrix} x(t) + \begin{bmatrix} e^{-t} \\ 0 \end{bmatrix} u(t)$$

One can readily show that the state transition matrix is

$$\mathbf{\Phi}(t;\tau) = \begin{bmatrix} e^{-(t-\tau)} & \frac{1}{2}\left(e^{t+\tau} - e^{-t+3\tau}\right) \\ 0 & e^{-(t-\tau)} \end{bmatrix}$$

This implies that

$$\mathbf{\Phi}(t;\tau)\mathbf{B}(\tau) = \begin{bmatrix} e^{-(t-\tau)} & \frac{1}{2}(e^{t+\tau} - e^{-t+3\tau}) \\ 0 & e^{-(t-\tau)} \end{bmatrix} \begin{bmatrix} e^{-\tau} \\ 0 \end{bmatrix} = \begin{bmatrix} e^{-t} \\ 0 \end{bmatrix}$$

So the reachability gramian is

$$\begin{aligned} \mathbf{W}_r(t_0, t_1) &= \int_{t_0}^{t_1} \mathbf{\Phi}(t_1;\tau)\mathbf{B}(\tau)\mathbf{B}^T(\tau)\mathbf{\Phi}^T(t_1;\tau)d\tau \\ &= \int_{t_0}^{t_1} \begin{bmatrix} e^{-t_1} \\ 0 \end{bmatrix} \begin{bmatrix} e^{-t_1} & 0 \end{bmatrix} d\tau \\ &= \begin{bmatrix} \tau e^{-2t_1} & 0 \\ 0 & 0 \end{bmatrix}\Bigg|_{t_0}^{t_1} = \begin{bmatrix} (t_1 - t_0)e^{-2t_1} & 0 \\ 0 & 0 \end{bmatrix} \end{aligned}$$

Clearly $\mathrm{rank}(\mathbf{W}_r(t_0, t_1)) = 1 < 2$ for any $t_0 < t_1$. So this system is not reachable for any $t_1$.

Even though $(\mathbf{A}(t), \mathbf{B}(t))$ is not a reachable pair, there are reachable states. In particular, the states in $\text{Range}(\mathbf{W}_r(t_0, t_1)) = \text{span}\left\{\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right\}$ are reachable. So if we consider any state $x_1 = \begin{bmatrix} \alpha \\ 0 \end{bmatrix}$ where $\alpha$ is any real number then $x_1$ must be reachable. We can use our earlier theorem to obtain a specific input transferring the state from the origin to this $x_1$. This is done by first solving the LAE

$$\mathbf{W}_r(t_0, t_1)\eta_1 = \begin{bmatrix} (t_1 - t_0)e^{-2t_1} & 0 \\ 0 & 0 \end{bmatrix} \eta_1 = \begin{bmatrix} \alpha \\ 0 \end{bmatrix} = x_1$$

which has the solution

$$\eta_1 = \begin{bmatrix} \frac{\alpha}{t_1 - t_0} e^{2t_1} \\ \beta \end{bmatrix}$$

where $\beta$ is any real number. Our earlier theorem therefore allows us to conclude that one control steering the state to $x_1$ will be

$$\begin{aligned} u(t) &= [\mathbf{\Phi}(t_1; t)\mathbf{B}(t)]^T \eta_1 \\ &= \begin{bmatrix} e^{-t_1} & 0 \end{bmatrix} \begin{bmatrix} \frac{\alpha}{t_1 - t_0} e^{2t_1} \\ \beta \end{bmatrix} = \frac{\alpha}{t_1 - t_0} e^{t_1} \end{aligned}$$

We can verify that this control input actually reaches the desired state by substituting $u$ back into the equation for $x_1$

$$\begin{aligned} x(t_1) &= \int_{t_0}^{t_1} \mathbf{\Phi}(t_1; \tau)\mathbf{B}(\tau)u(\tau)d\tau \\ &= \begin{bmatrix} e^{-t_1} \\ 0 \end{bmatrix} \frac{\alpha}{t_1 - t_0} e^{t_1}(t_1 - t_0) = \begin{bmatrix} \alpha \\ 0 \end{bmatrix} \end{aligned}$$

**Time Invariant Systems:** We now specialize the results above to continuous-time LTI systems

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

In this case the transition matrix may be written as $\mathbf{\Phi}(t;\tau) = e^{\mathbf{A}(t-\tau)}$. Because of time invariance, we take $t_0 = 0$ and $t_1 = T$ without loss of generality and the reachability gramian becomes

$$\mathbf{W}_r(0,T) = \int_0^T e^{(T-\tau)\mathbf{A}}\mathbf{B}\mathbf{B}^T e^{(T-\tau)\mathbf{A}^T} d\tau$$

When the system is time invariant, the range space of $\mathbf{W}_r(0,T)$ is independent of $T$. To show this, let us first assume that $\mathbf{B}$ is a vector so we are considering a scalar input $u$. Let us then consider $x_1 \in \mathrm{Range}(\mathbf{W}_r(0,T))$ for some $T > 0$. This would mean that there exists an input $u$ such that

$$\mathbf{L}[u;0,T] = \int_0^T e^{(T-\tau)\mathbf{A}}\mathbf{B}u(\tau)d\tau = x_1$$

Since $e^{\mathbf{A}t}$ is a power series, we can use the above equation to see that

$$x_1 = \sum_{k=0}^{\infty} \mathbf{A}^k\mathbf{B}\left(\int_0^T \frac{(T-\tau)^k}{k!}u(\tau)d\tau\right)$$

Invoking the Cayley-Hamilton theorem lets us express this infinite series as a finite series

$$x_1 = \sum_{k=0}^{n-1} \mathbf{A}^k\mathbf{B}\alpha_k(T)$$

where $\alpha_k : \mathbb{R} \to \mathbb{R}$ are functions of time. This last relation relation implies $x_1$ lies in the span of the collection of vectors $\{\mathbf{A}^k\mathbf{B}\}_{k=0}^{n-1}$. In particular, this subspace is the range space of a matrix whose columns are formed from these vectors,

$$\mathcal{C} = \begin{bmatrix} \mathbf{B} & \mathbf{AB} & \mathbf{A}^2\mathbf{B} & \cdots & \mathbf{A}^{n-1}\mathbf{B} \end{bmatrix}$$

This matrix is also called the pair's *controllability matrix*. So we can conclude that $x_1 \in \mathrm{Range}(\mathcal{C})$ and so $\mathrm{Range}(\mathbf{W}_r(0,T)) \subset \mathrm{Range}(\mathcal{C})$ for any $T > 0$.

Conversely, let us assume there exists $\eta_1 \in \mathbb{R}^m$ such that $\mathcal{C}\eta_1 = x_1$ and let us further assume that $x_1 \notin \mathrm{Range}(\mathbf{W}_r(0,T))$ for some $T > 0$. This

would imply that the null space of $\mathbf{W}_r(0,T)$ is nontrivial and so

$$\text{Range}(\mathbf{W}_r(0,T)) = (\ker(\mathbf{W}_r^T(0,cT))^\perp = (\ker(\mathbf{W}_r(0,T)))^\perp$$

where the last equality occurs because $\mathbf{W}_r$ is symmetric. So for any $w \in \text{Range}(\mathbf{W}_r)$ and $v \in \ker(\mathbf{W}_r)$ we can conclude $w^T v = 0$.

We then use the same argument leading to our earlier theorem to deduce there exists a vector $x_2 \in \ker(\mathbf{W}_r(0,T))$ such that $x_2^T x_1 \neq 0$. We then use the fact that $x_2^T \mathbf{W}_r(0,T)x_2 = 0$ to deduce that $x_2^T e^{(T-\tau)\mathbf{A}}\mathbf{B} = 0$ for all $0 \leq \tau \leq T$. By the Cayley-Hamilton theorem this would mean

$$0 = x_2^T \sum_{k=0}^{n-1} \mathbf{A}^k \mathbf{B} \alpha_k(T)$$

which would mean $x_2$ is orthogonal to $\text{Range}(\mathcal{C})$. This would mean that

$$x_2^T x_1 = x_2^T \mathcal{C} \eta_1 = 0$$

which contradicts our earlier observation that $x_2^T x_1 \neq 0$ if $x_1 \notin \text{Range}(\mathbf{W}_r)$. So by this contraction we can deduce $\text{Range}(\mathcal{C}) \subset \text{Range}(\mathbf{W}_r)$. Combining this with our earlier result we can see that $\text{Range}(\mathcal{C}) = \text{Range}(\mathbf{W}_r(0,T))$. Since the matrix $\mathcal{C}$ is independent of $T$, this means the subspace $\text{Range}(\mathbf{W}_r(0,T))$ is the same for all $T$. These observations can now be summarized in the following theorem.

THEOREM 20. *Consider the LTI system, $\dot{x}(t) = \mathbf{A}x(t) + \mathbf{B}u(t)$, then*

- $\text{Range}(\mathbf{W}_r(0,T)) = \text{Range}(\mathcal{C})$ *for all* $T > 0$
- *The reachable subspace is* $\mathcal{R} = \text{Range}(\mathcal{C})$
- *There exists an input* $u$

  $$u(t) = \mathbf{B}^T e^{\mathbf{A}^T(T-t)} \eta_1$$

  *transfers the state from the origin to* $x_1 \in \text{Range}(\mathcal{C})$ *by time* $T$ *for* $\eta_1$ *satisfying the LAE* $\mathcal{C}\eta_1 = x_1$.
- $(\mathbf{A}, \mathbf{B})$ *is a reachable/controllable pair if and only if* $\text{rank}(\mathcal{C}) = n$.

**Discrete-time Systems:** Consider the discrete-time LTI system

$$x(k+1) = \mathbf{A}x(k) + \mathbf{B}u(k)$$

Consider the problem of steering the state from $x_0$ to $x_1$ in finite time $K$. This means

$$\begin{aligned} x_1 &= \mathbf{A}^K x_0 + \sum_{i=0}^{K-1} \mathbf{A}^{K-(i+1)}\mathbf{B}u(i) \\ &= \mathbf{A}^K x_0 + \mathcal{C}_K \mathcal{U}_K \end{aligned}$$

where

$$\begin{aligned} \mathcal{C}_K &= \begin{bmatrix} \mathbf{B} & \mathbf{AB} & \cdots & \mathbf{A}^{K-1}\mathbf{B} \end{bmatrix} \\ \mathcal{U}_K &= \begin{bmatrix} u(K-1) \\ u(K-2) \\ \vdots \\ u(0) \end{bmatrix} \end{aligned}$$

The definitions for a state $x_1$ to be reachable or controllable at $K$ are identical to the definitions in the continuous-time case. To determine an input that steers a reachable state, however, can be determined directly from the the matrices of the system equations. Recall that if $x_1$ is reachable from the origin that

$$x_1 = \sum_{i=0}^{K-1} \mathbf{A}^{K-(i+1)}\mathbf{B}u(i) = \mathcal{C}_K \mathcal{U}_K$$

The last equation is a linear algebraic equation whose solution, $\mathcal{U}_K$, gives the desired control sequence. This simple argument can be summarized in the following theorem

THEOREM 21. *For a discrete time LTI system, the state $x_1$ is reachable from the origin in finite time if and only if $x_1 \in \mathrm{Range}(\mathcal{C})$. The reachable subspace of the system is $\mathcal{R}_r = \mathrm{Range}(\mathcal{C})$ and one input sequence that transfers the state from the origin to $x_1$ in $n$ steps is*

$$\mathcal{U}_n = \begin{bmatrix} u^T(n-1) & u^T(n-2) & \cdots & u^T(0) \end{bmatrix}^T$$

*where $\mathcal{U}_n$ satisfies $\mathcal{C}\mathcal{U}_n = x_1$.*

Now let us consider the problem of controlling the system state from $x_0$ to the origin in $K$ steps. We can see that this requires $x_0$ satisfy

$$-\mathbf{A}^K x_0 \in \text{Range}(\mathcal{C}_K)$$

One input that achieves this transfer is $\mathcal{U}_K$ that satisfies

$$-\mathbf{A}^K x_0 = \mathcal{C}_K \mathcal{U}_K$$

The fact that the left hand side of the equation passes our target through $\mathbf{A}^K$ means that controllability will not be equivalent to reachability. The reason for this is that $\mathbf{A}$ may not be invertible in a discrete-time system.

In particular, let us assume that $x$ is reachable. This would mean $x \in \text{Range}(\mathcal{C})$. But also note that if $x \in \text{Range}(\mathcal{C})$ then $\mathbf{A}^k x \in \text{Range}(\mathcal{C})$ for any $k \geq 0$. We refer to such subspaces as being $\mathbf{A}$-invariant. Since $x$ being reachable also implies $\mathbf{A}^k x \in \text{Range}(\mathcal{C})$ we can conclude that the preceding LAE, $-\mathbf{A}^K x_0 = \mathcal{C}_k \mathcal{U}_K$ has a solution and so $x$ is also controllable. Since our choice of $x$ was any state in the reachable subspace, we can conclude that if state $x$ is reachable it is also controllable.

But the converse relation need not be true. Controllability does not imply reachability for discrete-time systems. To verify this claim it suffices to find just one example. So consider

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

The system is controllable since any initial state $x_0$ can be returned to the origin in a single step using the input

$$u = -x_1(0) - x_2(0)$$

This assertion is verified by a direct computation

$$\begin{aligned} x_1 &= \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} x(0) + \begin{bmatrix} 1 \\ 0 \end{bmatrix} (-x_1(0) - x_2(0)) \\ &= \begin{bmatrix} x_1(0) + x_2(0) - x_1(0) - x_2(0) \\ 0 \end{bmatrix} = 0 \end{aligned}$$

This system, however, is not reachable as can be verified by simply finding the reachable subspace and noting it is not all of $\mathcal{R}^2$. In particular, we see that

$$\begin{aligned} x(1) &= \begin{bmatrix} 1 \\ 0 \end{bmatrix} u(0) = \begin{bmatrix} u(0) \\ 0 \end{bmatrix} \\ x(2) &= \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} x(1) + \begin{bmatrix} u(1) \\ 0 \end{bmatrix} = \begin{bmatrix} u(0) + u(1) \\ 0 \end{bmatrix} \\ \vdots &\quad \vdots \end{aligned}$$

and so we see that the reachable subspace is $\mathrm{span}(\left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right\}$. Since this is clearly not all of $\mathcal{R}^2$, it means this system has states that are controllable, but that are not reachable.

**Example:** Consider the discrete-time LTI system

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

(1) Is the system reachable? Determine the reachability subspace. Consider the target state $x_1 = \begin{bmatrix} -2 \\ 2 \\ 4 \end{bmatrix}$ and find an input that reaches $x_1$ from the origin. Verify that your inputs actually reach the desired $x_1$.

(2) Is the system controllable? Determine the set of states that can be controlled to the origin in finite time. Characterize all inputs that

drive the system to the origin in two steps. Verify that your input actually drives the system state to the origin.

For the reachability problem, we first find the controllability matrix, $\mathcal{C}$,

$$\mathcal{C} = \begin{bmatrix} \mathbf{B} & \mathbf{AB} & \mathbf{A^2B} \end{bmatrix} = \begin{bmatrix} -1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

The matrix clearly has a rank of 2 and so this system is not reachable. The reachable subspace is

$$\mathcal{R}_r = \text{span}\left\{ \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right\}$$

The desired target state is

$$x_1 = \begin{bmatrix} -2 \\ 2 \\ 4 \end{bmatrix} = -2 \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix} + 4 \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

So $x_1$ is in the reachable subspace. An input $\mathcal{U} = [u(2), u(1), u(0)]^T$ that satisfies

$$\mathcal{CU} = \begin{bmatrix} u(2) \\ u(1) \\ u(0) \end{bmatrix} = \begin{bmatrix} -2 \\ 2 \\ 4 \end{bmatrix}$$

which has the solution

$$\begin{aligned} \mathcal{U} &= \begin{bmatrix} 2 \\ 2 \\ 0 \end{bmatrix} + \ker(\mathcal{C}) \\ &= \begin{bmatrix} 2 \\ 2 \\ 0 \end{bmatrix} + \alpha \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 2 \\ 2+\alpha \\ -\alpha \end{bmatrix} \end{aligned}$$

where $\alpha$ is any real number. We can verify this input works by computing

$$\begin{aligned} x(1) &= \mathbf{A}0 + \mathbf{B}(-\alpha) = \begin{bmatrix} \alpha \\ -\alpha \\ -\alpha \end{bmatrix} \\ x(2) &= \mathbf{A}x(1) + \mathbf{B}(2+\alpha) = \begin{bmatrix} -\alpha - 2 \\ \alpha + 2 \\ 2 \end{bmatrix} \\ x(3) &= \mathbf{A}x(2) + \mathbf{B}2 = \begin{bmatrix} -2 \\ 2 \\ 4 \end{bmatrix} \end{aligned}$$

Now let us turn to the controllability part of the problem. All states that can be driven to the origin in $K$ steps must satisfy

$$\mathbf{A}^K x_0 \in \text{Range}(\mathcal{C}_K)$$

In our case $K = 2$ and $\mathbf{A}^2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$. So for any $x_0 = \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix} \in \mathbb{R}^3$, we can see that $\mathbf{A}^2 x_0 = \begin{bmatrix} 0 \\ 0 \\ \gamma \end{bmatrix}$. The range space of $\mathcal{C}_2$ is

$$\text{Range}(\mathcal{C}_2) = \text{Range}\left(\begin{bmatrix} -1 & 0 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}\right) = \text{span}\left\{\begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}\right\}$$

So for any $x_0 \in \mathbb{R}^3$, we have $\mathbf{A}^2 x_0 \in \text{Range}(\mathcal{C}_2)$, so every state in $\mathbb{R}^3$ is controllable in two steps and the system is controllable. The set of all inputs that drive $x_0$ to the origin in two steps is

$$-\mathbf{A}^2 x_0 = -\begin{bmatrix} 0 \\ 0 \\ \gamma \end{bmatrix} = \mathcal{C}_2 \begin{bmatrix} u(1) \\ u(0) \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} u(1) \\ u(0) \end{bmatrix}$$

The null space of $\mathcal{C}_2$ is trivial so there is a unique input that satisfies the LAE. That input is

$$\mathcal{U} = \begin{bmatrix} u(1) \\ u(0) \end{bmatrix} = \begin{bmatrix} 0 \\ -\gamma \end{bmatrix}$$

We verify the correctness of this control by computing the states

$$\begin{aligned} x(1) &= \begin{bmatrix} 1 & 1 & 0 \\ -1 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix} + \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix} (-\gamma) = \begin{bmatrix} \alpha+\beta+\gamma \\ -\alpha-\beta-\gamma \\ 0 \end{bmatrix} \\ x(2) &= \begin{bmatrix} 1 & 1 & 0 \\ -1 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix} x(1) + \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix} 0 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \end{aligned}$$

## 3. Observability and Constructibility Definitions

Observability is a finite-time property of a linear state-based system that is "dual" to reachability. In particular observability means that one can deduce the system's initial state, $x_0$, at time $t_0$ from a finite duration of inputs and outputs after time $t_0$. Constructibility is similar to controllability in that it reverses time and means that one can deduce the system's state, $x_1$ at time $t_1$ using finite duration inputs and outputs observed before time $t_1$.

Observability can be informally explained using discrete-time LTI system

$$\begin{aligned} x(k+1) &= \mathbf{A}x(k) + \mathbf{B}u(k) \\ y(k) &= \mathbf{C}x(k) + \mathbf{D}u(k) \end{aligned}$$

In this case we know the system output is

$$y(k) = \mathbf{C}\mathbf{A}^k x(0) + \sum_{i=0}^{k-1} \mathbf{C}\mathbf{A}^{k-(i+1)}\mathbf{B}u(i) + \mathbf{D}u(k)$$

for $k \geq 0$. This implies that

$$\begin{aligned}\widetilde{y}(k) &\stackrel{\text{def}}{=} y(k) - \left\{ \sum_{i=0}^{k-1} \mathbf{C}\mathbf{A}^{k-(i+1)}\mathbf{B}u(i) + \mathbf{D}u(k) \right\} \\ &= \mathcal{C}\mathcal{A}^k x_0\end{aligned}$$

for $k \geq 0$. Note that $\widetilde{y}(k)$ is known for $0 \leq k \leq K$, so the initial state is obtained as a solution to the linear algebraic equation

$$\widetilde{Y}_{0,K-1} = \mathcal{O}_K x_0$$

where

$$\mathcal{O}_K = \begin{bmatrix} \mathbf{C} \\ \mathbf{CA} \\ \vdots \\ \mathbf{CA}^{K-1} \end{bmatrix}, \quad \widetilde{Y}_{0,n-1} = \begin{bmatrix} \widetilde{y}(0) \\ \widetilde{y}(1) \\ \vdots \\ \widetilde{y}(K-1) \end{bmatrix}$$

Whether or not we can determine $x_0$ from a given set of $K$ measured inputs, $u$ and outputs, $y$, will depend on whether the preceding linear algebraic equation has a unique solution. This is the basic approach adopted when extending these ideas to continuous-time and time-varying linear systems.

We now formalize our definition of observability for continuous-time LTV systems

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

The output $y(t)$ is

$$y(t) = \mathbf{C}(t)\mathbf{\Phi}(t;t_0)x(t_0) + \int_{t_0}^{t} \mathbf{C}(t)\mathbf{\Phi}(t\tau)\mathbf{B}(\tau)u(\tau)d\tau + \mathbf{D}(t)u(t)$$

where $\mathbf{\Phi}(t;\tau)$ is the state transition matrix. We can rewrite the above equation as

$$\widetilde{y}(t) = \mathbf{C}(t)\mathbf{\Phi}(t;t_0)x_0$$

where $x_0 = x(t_0)$ and where

$$\widetilde{y}(t) \stackrel{\text{def}}{=} y(t) - \left\{ \int_{t_0}^{t} \mathbf{C}(t)\mathbf{\Phi}(t;\tau)\mathbf{B}(\tau)u(\tau)d\tau + \mathbf{D}(t)u(t) \right\}$$

A state $x \in \mathbb{R}^n$ is *unobservable* at time $t_0$ if the zero-input (natural) response of the system is zero for all $t \geq t_0$. In other words, $x$ is unobservable at $t_0$ if and only if

$$\mathbf{C}(t)\mathbf{\Phi}(t;t_0)x = 0, \quad \text{for all } t \geq t_0$$

The *unobservable subspace* at $t_0$ is denoted as $\mathcal{R}_o^{t_0}$ and consists of all states that are unobservable at $t_0$. We say the system is completely observable at $t_0$ if and only if the only unobservable state is the origin. In other words, the system is *observable* if and only i $\mathcal{R}_o^{t_0} = \{0\}$. If the system is observable, then we refer to $(\mathbf{A}(t), \mathbf{C}(t))$ as an *observable pair*.

Observability uses future output/inputs to determine the initial state at time $t_0$. Constructibility uses past inputs/outputs prior to the initial time, $t_0$, to determine the initial state, $x_0$. So a state $x$ is *unconstructible* at time $t_1$ if and only if for all $t \geq t_1$, the zero input (natural response) of the system is zero

$$\mathbf{C}(t)\mathbf{\Phi}(t;t_1)x = 0, \quad \text{for all } t \leq t_1$$

The unconstructible at $t_1$ subspace is denoted as $\mathcal{R}_{cn}^{t_1}$ and consists of all states that are unconstructible at time $t_1$. The system is *constructible* at $t_1$ if and only if the only state that is unconstructible at $t_1$ is the origin (i.e. $\mathcal{R}_{cn}^{t_1} = \{0\}$). If the system is constructible then we refer to $(\mathbf{A}(t), \mathbf{C}(t))$ as a *constructible pair*.

As before, observability and constructibility are equivalent for continuous-time LTV systems and may not be equivalent for discrete-time systems. To prove the equivalence of these concepts for continuous-time systems, we define the *observability gramian* as a matrix-valued function, $\mathbf{W}_o : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^{n \times n}$ that takes values

$$\mathbf{W}_o(t_0, t_1) = \int_{t_0}^{t_1} \mathbf{\Phi}^T(\tau;t_0)\mathbf{C}^T(\tau)\mathbf{C}(\tau)\mathbf{\Phi}(\tau;t_0)d\tau$$

Note that $\mathbf{W}_o(t_0, t_1)$ is symmetric and positive semidefinite for all $t_1 > t_0$. So if $x$ is unobservable at $t_0$, then

$$\mathbf{C}(t)\mathbf{\Phi}(t; t_0)x = 0, \quad \text{for all } t \geq t_0$$

This means that

$$\mathbf{W}_o(t_0, t_1)x \;=\; \int_{t_0}^{t_1} \mathbf{\Phi}^T(\tau; t_0)\mathbf{C}^T(\tau)\mathbf{C}(\tau)\mathbf{\Phi}(\tau; t_0)x d\tau = 0$$

for all $t_1 \geq t_0$. The last equality holds because $\mathbf{C}\mathbf{\Phi}x = 0$ and so we can conclude

$$x \in \ker(\mathbf{W}_o(t_0, t_1))$$

This implies that $x$ being in the null space of the observability gramian is necessary for $x$ being unobservable at $t_0$.

Conversely, let $x$ be in the null space of $\mathbf{W}_o(t_0, t_1)$ for all $t_1 \geq t_0$. This means that

$$0 \;=\; x^T\mathbf{W}_o(t_0, t_1)x = \int_{t_0}^{t_1} \left|\mathbf{C}(\tau)\mathbf{\Phi}(\tau; t_0)x\right|^2 d\tau$$

for all $t_1 \geq t_0$. This can only occur if $|\mathcal{C}(\tau)\mathbf{\Phi}(\tau; t_0)x| = 0$ for all $\tau > t_o$ which is only true if $\mathbf{C}(\tau)\mathbf{\Phi}(\tau; t_0)x = 0$ for all $\tau > t_0$, or rather that $x$ is unobservable at $t_0$. So $x$ being in the null space of the observability gramian is also sufficient for $x$ being unobservable.

The preceding argument says that $x$ is unobservable at $t_0$ if and only if $x$ lies in the null space of $\mathbf{W}_o(t_0, t_1)$ for all $t_1 > t_0$. If $x$ is observable at $t_0$, then there is a $t_1 > t_0$ such that the null space of $\mathbf{W}_o(t_0, t_1)$ is trivial or rather than $\mathrm{rank}(\mathbf{W}_o(t_0, t_1)) = n$ for some $t_1 > t_0$. This condition can be readily verified by checking the determinant of $\mathbf{W}_o(t_0, t_1)$.

We can also define a constructibility gramian

$$\mathbf{W}_{\mathrm{cn}}(t_0, t_1) = \int_{t_0}^{t_1} \mathbf{\Phi}^T(\tau; t_1)\mathbf{C}^T(\tau)\mathbf{C}(\tau)\mathbf{\Phi}(\tau, t_1)d\tau$$

A similar argument can be used to show that $x$ is unconstructible at $t_1$ if and only if $x$ lies in the kernel of $\mathbf{W}_{\mathrm{cn}}(t_0, t_1)$ for all $t_0 < t_1$. In a similar way,

we can assert that $x$ is constructible if there is $t_0 < t_1$ such that $\mathbf{W}_{\text{cn}}(t_0, t_1)$ has full rank; a condition that can be checked using the determinant of the constructibility gramian.

Note that

$$\mathbf{W}_o(t_1, t_0) = \mathbf{\Phi}^T(t_1; t_0)\mathbf{W}_{\text{cn}}(t_0, t_1)\mathbf{\Phi}(t_1; t_0)$$

Since the transition matrix for a continuous time system is nonsingular, we can conclude that

$$\text{rank}(\mathbf{W}_o(t_1, t_0)) = \text{rank}(\mathbf{W}_{\text{cn}}(t_0, t_1))$$

thereby establishing that $\mathbf{W}_o$ has rank $n$ if and only if $\mathbf{W}_{\text{cn}}$ has rank $n$, which implies observability and constructibility are equivalent in continuous time systems. This result is summarized in the following theorem.

THEOREM 22. *Consider the LTV system $\dot{x}(t) = \mathbf{A}(t)x(t)$ with output $y(t) = \mathbf{C}(t)x(t)$. The state $x$ is unconstructible at $t_1$ if and only if it is unobservable at $t_0$. The system is completely observable if and only if it completely constructible.*

## 4. Conditions for Observability/Constructibility

This section develops conditions that can be checked to see if a state or system is observable or constructible. We will confine our attention to LTI systems. Let us first consider the continuous-time LTI system

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

The output will be

$$y(t) = \mathbf{C}e^{\mathbf{A}t}x_0 + \int_0^t \mathbf{C}e^{\mathbf{A}(t-\tau)}\mathbf{B}u(\tau)d\tau + \mathbf{D}u(t)$$

for $t \geq 0$. As before we rewrite the output equation as

$$\begin{aligned} \widehat{y}(t) &= y(t) - \left\{ \int_0^t \mathbf{C}e^{\mathbf{A}(t-\tau)}\mathbf{B}u(\tau)d\tau + \mathbf{D}u(t) \right\} \\ &= \mathbf{C}e^{\mathbf{A}t}x_0 \end{aligned}$$

By definition, this state $x_0$ is uobservable if the zero-input response of the system is zero for all $t \geq 0$. In other words, $x$ is unobservable if

$$\mathbf{C}e^{\mathbf{A}t}x_0 = 0$$

for all $t \geq 0$. This set of unobservable states is denoted as $\mathcal{R}_{\overline{o}}$ and $(\mathbf{A}, \mathbf{C})$ is an observable pair if $\mathcal{R}_{\overline{o}} = \{0\}$ or rather if the observability gramian

$$\mathbf{W}_o(0,t) = \int_0^T e^{\mathbf{A}^T\tau}\mathbf{C}^T\mathbf{C}e^{\mathbf{A}\tau}d\tau$$

has full rank for any $t \geq 0$. Evaluation of the observability gramian for all time $t$ may be complicated so we seek a simpler condition that is not dependent on $t$. In particular, define the *observability matrix* as

$$\mathcal{O} = \begin{bmatrix} \mathbf{C} \\ \mathbf{CA} \\ \vdots \\ \mathbf{CA}^{n-1} \end{bmatrix}$$

We will show that the null space of $\mathbf{W}_o$ is independent of $t$ and equals $\ker(\mathcal{O})$.

To prove this assertion, first let $x \in \ker(\mathcal{O})$ so that $\mathcal{O}x = 0$. From the definition of $\mathcal{O}$ this means $\mathbf{CA}^k x = 0$ for all $0 \leq k \leq n-1$. Taking the series expansion for $e^{\mathbf{A}t}$, we rewrite the zero-input response as

$$\mathbf{C}e^{\mathbf{A}t}x = \mathbf{C}\left\{ \sum_{k=0}^{\infty} \frac{t^k}{k!}\mathbf{A}^k \right\} x$$

for all $t \geq 0$. Now look at the gramian

$$\begin{aligned}
\mathbf{W}_o(0,T)x &= \int_0^T e^{\mathbf{A}^T\tau}\mathbf{C}^T\mathbf{C}e^{\mathbf{A}\tau}x d\tau \\
&= \int_0^T e^{\mathbf{A}\tau}\mathbf{C}^T\mathbf{C}\left\{\sum_{k=0}^{\infty}\frac{\tau^k}{k!}\mathbf{A}^k\right\}x d\tau \\
&= \sum_{k=0}^{\infty}\left\{\int_0^T e^{\mathbf{A}^T\tau}\mathbf{C}^T\frac{\tau^k}{k!}d\tau\right\}\mathbf{C}\mathbf{A}^k x \\
&= 0
\end{aligned}$$

The last line occurs because we assumed $\mathbf{C}\mathbf{A}^k x = 0$. So we can conclude that $x \in \ker(\mathbf{W}_o(0,T))$ for all $T > 0$ which means $\ker(\mathcal{O}) \subset \ker(\mathbf{W}_o(0,T))$ for all $T$.

Conversely, let $x \in \ker(\mathbf{W}_o(0,T))$ for some $T > 0$ and examine the quadratic form

$$x^T\mathbf{W}_o(0,T)x = \int_0^T \left|\mathbf{C}e^{\mathbf{A}t}x\right|^2 dt = 0$$

with the last equality holding because $x$ is in the gramian's null space. This integral can only be true if $\left|\mathbf{C}e^{\mathbf{A}t}x\right| = 0$, which implies $\mathbf{C}e^{\mathbf{A}t}x = 0$ for every $t \in [0,T]$. Taking the $k$th derivative of $\mathbf{C}e^{\mathbf{A}t}x$ with respect to $t$ and evaluating at $t = 0$ yields,

$$\mathbf{C}x = \mathbf{C}\mathbf{A}x = \cdots = \mathbf{C}\mathbf{A}^k x = 0$$

for all $k$. This implies that $\mathcal{O}x = 0$ and so $\ker(\mathbf{W}_o(0,T)) \subset \ker(\mathcal{O})$. Taking this result along with our earlier deduction that $\ker(\mathcal{O}) \subset \ker(\mathbf{W}_o)$ leads to the following theorem.

THEOREM 23. *For the continuous-time LTI system,* $\ker(\mathcal{O}) = \ker(\mathbf{W}_o(0,T))$ *for all* $T \geq 0$.

Because $\ker(\mathcal{O}) = \ker(\mathbf{W}_o(0,T))$ for all $T \geq 0$, we can readily see that a state $x$ is unobservable if and only if $x \in \ker(\mathcal{O})$. Moreover, we can also see that $(\mathbf{A},\mathbf{C})$ is an observable pair if and only if $\mathcal{R}_{\overline{o}} = \{0\}$. This means that $\operatorname{rank}(\mathcal{O}) = n$ is a necessary and sufficient condition for $(\mathbf{A},\mathbf{C})$ to be

an observable pair. Assuming that the system is observable, then one can determine $x_0$ as follows. Take the equation

$$\widetilde{y}(\tau) = \mathbf{C}e^{\mathbf{A}\tau}x_0$$

Pre-multiply by $e^{\mathbf{A}^T\tau}\mathbf{C}^T$ and integrating from $0$ to $T$ yields

$$\int_0^T e^{\mathbf{A}^T\tau}\mathbf{C}^T\widetilde{y}(\tau)d\tau \;\; = \;\; \int_0^T e^{\mathbf{A}^T\tau}\mathbf{C}^T\mathbf{C}e^{\mathbf{A}\tau}d\tau x_0 = \mathbf{W}_o(0,T)x_0$$

If the system is observable then $\mathbf{W}_o(0,T)$ is invertible and so the we can solve the preceding linear equation to get

$$x_0 = \mathbf{W}_o^{-1}(0,T)\left\{\int_0^T e^{\mathbf{A}^T\tau}\mathbf{C}^T\widetilde{y}(\tau)d\tau\right\}$$

We can summarize our results in the following theorem.

THEOREM 24. *The LTI system is observable if and only if* $\mathrm{rank}(\mathcal{O}) = n$. *The initial state at time* $0$ *may then be obtained from observed inputs,* $u$, *and outputs* $y$ *over the interval* $[0,T]$ *as*

$$x_0 = \mathbf{W}_o^{-1}(0,T)\left\{\int_0^T e^{\mathbf{A}^T\tau}\mathbf{C}^T\widetilde{C}^T\widetilde{y}(\tau)d\tau\right\}$$

Note that in general $x_0$ is not determined directly using this formula. The reason is that for small $T$, the observability gramian's inverse may be very sensitive to small perturbations or noise in the data. In practice we use recursive estimation algorithms to estimate the initial state in a manner that minimizes the estimate's covariance. Such "estimators" will be discussed in the next chapter.

For discrete-time LTI systems, observability and constructibility are not equivalent. To see this, note that the system's output is

$$y(k) \;\; = \;\; \mathbf{C}\mathbf{A}^k x(0) + \sum_{i=0}^{k-1}\mathbf{C}\mathbf{A}^{k-(i+1)}\mathbf{B}u(i) + \mathbf{D}u(k)$$

for $k > 0$. We rewrite this as

$$\begin{aligned}\widetilde{y}(k) &\stackrel{\text{def}}{=} y(k) - \left(\sum_{i=0}^{k-1} \mathbf{C}\mathbf{A}^{k-(i+1)}\mathbf{B}u(i) + \mathbf{D}u(k)\right) \\ &= \mathbf{C}\mathbf{A}^k x_0\end{aligned}$$

For discrete time systems, we say that state $x$ is unobservable at $0$ if and only if

$$\mathbf{C}\mathbf{A}^k x = 0$$

for all $k \geq 0$. This is analogous to how we defined unobservable states in continuous time. We can, readily show that $x$ is unobservable if and only if it lies in the null space of the observability matrix. If that observability matrix has full rank, then we know the system is completely observable. To get $x_0$, note that

$$y(k) = \mathbf{C}\mathbf{A}^k x_0 + \sum_{i=0}^{k-1} \mathbf{C}\mathbf{A}^{k-(i+1)}\mathbf{B}u(i) + \mathbf{D}u(k)$$

which for $k = 0, 1, \ldots, n-1$ leads to the following system of linear equations

$$\begin{aligned} y(0) - \mathbf{D}u(0) &= \mathbf{C}x_0 \\ y(1) - \mathbf{C}\mathbf{B}u(0) - \mathbf{D}u(1) &= \mathbf{C}\mathbf{A}x_0 \\ y(2) - \mathbf{C}\mathbf{A}\mathbf{B}u(0) - \mathbf{C}\mathbf{B}u(1) - \mathbf{D}u(2) &= \mathbf{C}\mathbf{A}^2 x_0 \\ &\vdots \\ y(n-1) - \sum_{i=0}^{n-2} \mathbf{C}\mathbf{A}^{n-(i+1)}u(i) - \mathbf{D}u(n-1) &= \mathbf{C}\mathbf{A}^{n-1}x_0 \end{aligned}$$

which in matrix vector form is

$$\begin{aligned} \begin{bmatrix} \mathbf{C} \\ \mathbf{C}\mathbf{A} \\ \vdots \\ \mathbf{C}\mathbf{A}^{n-1} \end{bmatrix} x_0 &= \begin{bmatrix} y(0) \\ y(1) \\ \vdots \\ y(n-1) \end{bmatrix} - \begin{bmatrix} \mathbf{D} & 0 & \cdots & 0 & 0 \\ \mathbf{C}\mathbf{B} & \mathbf{D} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{C}\mathbf{A}^{n-2}\mathbf{B} & \mathbf{C}\mathbf{A}^{n-3}\mathbf{B} & \cdots & \mathbf{D} & 0 \\ \mathbf{C}\mathbf{A}^{n-1}\mathbf{B} & \mathbf{C}\mathbf{A}^{n-2}\mathbf{B} & \cdots & \mathbf{C}\mathbf{B} & \mathbf{D} \end{bmatrix} \begin{bmatrix} u(0) \\ u(1) \\ \vdots \\ u(n-2) \\ u(n-1) \end{bmatrix} \\ \mathbf{O}x_0 &= [\mathbf{Y}_{0,n-1} - \mathbf{M}_n \mathbf{U}_{0,n-1}] \end{aligned}$$

The matrix $\mathbf{M}_n$ is sometimes called a matrix of Markov parameters.

The notion of constructibility will be defined differently for discrete-time LTI systems than it was for continuous-time systems. This difference comes from the fact that the discrete-time system's transition matrix may be singular. For discrete-time systems, we say that $x$ is unconstructible if and only if for all $k \geq 0$ there exists a nonzero $\widehat{x} \in \mathbb{R}^n$ such that

$$x = \mathbf{A}^k \widehat{x} \quad \textbf{and} \quad \mathbf{C}\widehat{x} = 0$$

If $\mathbf{A}$ is nonsingular, then this is equivalent to saying that $\mathbf{C}\mathbf{A}^{-k}x = 0$, which corresponds to our definition of constructibility for continuous-time systems. The preceding definition essentially says there is a "prior" state, $\widehat{x}$ that reaches $x$ in a finite number of steps and that this prior state is indistinguishable from $0$. Results relating observability and constructibility in discrete-time are similar to results regarding reachability and controllability. Our main result is formalized in the following theorem

THEOREM 25. *For the discrete-time LTI system we have*

- *If the state $x$ is unconstructible, then $x$ is also unobservable*
- $\mathcal{R}_{\overline{\text{cn}}} \subset \mathcal{R}_{\overline{o}}$
- *If the system is observable, then it is also constructible*

Let us first verify the first assertion and assume that $x$ is unconstructible, so for every $k \geq 0$ there exists $\widehat{x}$ such that

$$x = \mathbf{A}^k \widehat{x}, \quad \mathbf{C}\widehat{x} = 0$$

Premultiply by $\mathbf{C}$ to get

$$\mathbf{C}x = \mathbf{C}\mathbf{A}^k \widehat{x}$$

for every $k \geq 0¿$. Note that $\mathbf{C}x = 0$. Therefore $\mathbf{C}\mathbf{A}^k \widehat{x} = 0$ for all $k$, which means $\widehat{x} \in \ker(\mathcal{O}$. The null space of $\mathcal{O}$ can be shown to be $\mathbf{A}$-invariant, so we also know that $\mathbf{A}\widehat{x} \in \ker(\mathcal{O})$, which would mean that $x$ is unobservable. So $x$ being unconstructible implies $x$ is unobservable. The remaining assertions are immediate consequences of this fact.

## 5. Standard Forms for Uncontrollable/Unobservable LTI Systems

The preceding sections provided necessary and sufficient conditions for a system realization, $\mathbf{G} \stackrel{s}{=} \left[\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right]$ to be controllable/reachable or observable/constructible. But clearly not all system realizations of interest to us may have such completely controllable/observable realizations. One example is shown below in Fig. 1. This figure shows an airfoil in a wind tunnel and the flows around that foil. The "state" space is obtained by gridding the 2-dimensional space around the foil and the state in each grid element denotes the type of flow in that grid cell. What we see is the formation of vortices in the cells containing air/wing interfaces and we would like to control the surface of the wing to reduce the size of these vortices since they create drag. The impact of a control surface on the wing, however, will not effect the states of the flows in all grid cells. In particular, those grid cells next to the wing and those cells after the wing containing vortices can be controlled by the control surfaces. But those grid cells far from the wing/air interface will be minimally impacted by our controls, if at all. So this is a partially controllable system since our control cannot impact all grid cells equally. We are not only concerned with whether we can control a distant state, but also the amount to which a controllable state can be effected. Moreover, we note that because of our gridding of the 2-d surface, the number of states is extremely large. Since many of these states are not effected by the control, then perhaps we can develop a *reduced order* realization of the wing's air flows by neglecting all those grid cells that are minimally impacted by the control surface. The utility of such reduced order realizations is that they are easier to work with computationally since they have fewer states. This example clearly shows the importance of the controllability/reachability concept. This section examines methods for characterizing partially reachable/observable systems. Later sections examine methods for model reduction based on these characterizations.

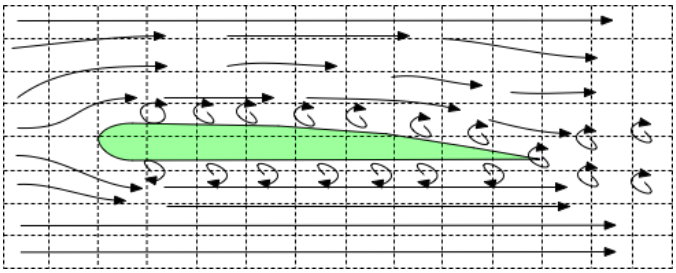


FIGURE 1. Aerodynamical Flows Example

Based on the preceding example, we can see that a realization's reachable subspace provides a basis for decomposing the original realization into controllable and uncontrollable subsystems. Such decompositions then provide the basis for useful schemes for model reduction. Unobservable subspaces provide a similar decomposition in terms observable and unobservable subsystems.

Let us begin by considering reachable decompositions for LTI continuous or discrete-time systems. If $(\mathbf{A}, \mathbf{B})$ is not a controllable pair, then we will show that we can decompose the system into controllable and uncontrollable subsystems through a similarity transformation. In particular, let $\text{rank}(\mathcal{C}) = n_r < n$ and let

$$\{v_1, v_2, \ldots, v_{n_r}\}$$

be a basis for the reachable subspace, $\mathcal{R}_r$. We will introduce the following nonsingular matrix

$$\mathbf{Q} = \begin{bmatrix} v_1 & v_2 & \cdots & v_{n_r} & \mathbf{Q}_{n-n_r} \end{bmatrix} \tag{39}$$

where $\mathbf{Q}_{n-n_r}$ is an $n \times (n-n_r)$ matrix whose linearly independent columns are chosen to ensure $\mathbf{Q}$ is nonsingular.

We know that the reachable subspace, $\mathcal{R}_r$ is $\mathbf{A}$-invariant and so $\mathbf{A}v_i \in \mathcal{R}_r$ for $i = 1, 2, \ldots, n_r$. This means that the columns of the matrix

$$\begin{bmatrix} \mathbf{A}v_1 & \cdots & \mathbf{A}v_{n_r} \end{bmatrix}$$

can be written as a linear combination of $v_1, \ldots, v_{n_r}$ and so

$$\mathbf{A}\begin{bmatrix} v_1 & \cdots & v_{n_r} \end{bmatrix} = \begin{bmatrix} v_1 & \cdots & v_{n_r} & \mathbf{Q}_{n-n_r} \end{bmatrix}\begin{bmatrix} \mathbf{A}_1 \\ \mathbf{0} \end{bmatrix}$$

where $\mathbf{A}_1$ is some $n_r \times n_r$ matrix. Because the columns of $\mathbf{B}$ are in $\mathrm{Range}(\mathcal{C}) = \mathcal{R}_r$, it should also be apparent that

$$\mathbf{B} = \begin{bmatrix} v_1 & \cdots & v_{n_r} & \mathbf{Q}_{n-n_r} \end{bmatrix}\begin{bmatrix} \mathbf{B}_1 \\ \mathbf{0} \end{bmatrix}$$

where $\mathbf{B}_1$ is an appropriately dimensioned matrix. This means that we can write $\mathbf{AQ}$ and $\mathbf{B}$ as

$$\begin{aligned} \mathbf{AQ} &= \mathbf{A}\begin{bmatrix} v_1 & \cdots & v_{n_r} & \mathbf{Q}_{n-n_r} \end{bmatrix} \\ &= \begin{bmatrix} v_1 & \cdots & v_{n_r} & \mathbf{Q}_{n-n_r} \end{bmatrix}\begin{bmatrix} \mathbf{A}_1 & \mathbf{A}_{12} \\ \mathbf{0} & \mathbf{A}_2 \end{bmatrix} = \mathbf{Q}\widehat{\mathbf{A}} \\ \mathbf{B} &= \begin{bmatrix} v_1 & \cdots & v_{n_r} & \mathbf{Q}_{n-n_r} \end{bmatrix}\begin{bmatrix} \mathbf{B}_1 \\ \mathbf{0} \end{bmatrix} = \widehat{\mathbf{B}}\mathbf{Q} \end{aligned}$$

These last equations imply that $(\widehat{\mathbf{A}}, \widehat{\mathbf{B}})$ is also a state space realization for the original system $(\mathbf{A}, \mathbf{B})$ where

$$\widehat{\mathbf{A}} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{A}_{12} \\ \mathbf{0} & \mathbf{A}_2 \end{bmatrix}, \quad \widehat{\mathbf{B}} = \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{0} \end{bmatrix}$$

The matrices $(\mathbf{A}_1, \mathbf{B}_1)$ are special because we can show they form a controllable pair. This assertion can be verified by computing the controllability matrix for $(\widehat{\mathbf{A}}, \widehat{\mathbf{B}})$,

$$\widehat{\mathcal{C}} = \mathbf{Q}^{-1}\mathcal{C} = \begin{bmatrix} \mathbf{B}_1 & \mathbf{A}_1\mathbf{B}_1 & \cdots & \mathbf{A}_1^{n-1}\mathbf{B}_1 \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

We know by assumption that $\mathrm{rank}(\mathcal{C}) = n_r < n$ and since $\mathbf{Q}$ is nonsingular this means that $\mathrm{rank}(\widehat{\mathcal{C}}) = n_r$. From the form we computed for $\widehat{\mathcal{C}}$, it should be apparent that

$$\mathrm{rank}\begin{bmatrix} \mathbf{B}_1 & \mathbf{A}_1\mathbf{B}_1 & \cdots & \mathbf{A}_1^{n-1}\mathbf{B}_1 \end{bmatrix} = n_r$$

which means that $(\mathbf{A}_1, \mathbf{B}_1)$ is a controllable pair. The preceding discussion can be summarized in the following theorem.

THEOREM 26. *If* $(\mathbf{A}, \mathbf{B})$ *is not a controllable pair, then there exists a nonsingular matrix* $\mathbf{Q}$ *such that*

$$\widehat{\mathbf{A}} = \mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{A}_{12} \\ \mathbf{0} & \mathbf{A}_2 \end{bmatrix}, \quad \widehat{\mathbf{B}} = \mathbf{Q}^{-1}\mathbf{B} = \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{0} \end{bmatrix}$$

*where* $(\mathbf{A}_1, \mathbf{B}_1)$ *is a controllable pair.*

The preceding theorem provides the basis for decomposing an uncontrollable state space realization $\mathbf{G} \stackrel{s}{=} \left[\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right]$ into controllable and uncontrollable subsystems. If we define a new state $\widehat{x} = \mathbf{Q}^{-1}x$ we get

$$\begin{aligned} \dot{\widehat{x}} = \begin{bmatrix} \dot{\widehat{x}}_1 \\ \dot{\widehat{x}}_2 \end{bmatrix} &= \begin{bmatrix} \mathbf{A}_1 & \mathbf{A}_{12} \\ \mathbf{0} & \mathbf{A}_2 \end{bmatrix} \begin{bmatrix} \widehat{x}_1 \\ \widehat{x}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{0} \end{bmatrix} u \\ y &= \begin{bmatrix} \mathbf{C}_1 & \mathbf{C}_2 \end{bmatrix} \begin{bmatrix} \widehat{x}_1 \\ \widehat{x}_2 \end{bmatrix} + \mathbf{D}u \end{aligned}$$

where $(\mathbf{A}_1, \mathbf{B}_1)$ is a controllable pair. We call this the *standard form* for uncontrollable LTI systems. The $n_r$ eigenvalues of $\mathbf{A}_1$ and their corresponding eigenvectors are called *controllable* eigenvalues/eigenvectors (modes) of the pair $(\mathbf{A}, \mathbf{B})$. The $n - n_r$ eigenvalues of $\mathbf{A}_2$ are the uncontrollable eigenvalues (modes) of the system.

**Example:** Consider the system

$$\dot{x} = \begin{bmatrix} 0 & -1 & 1 \\ 1 & -2 & 1 \\ 0 & 1 & -1 \end{bmatrix} x + \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} u$$

The controllability matrix is

$$\mathcal{C} = \left[\begin{array}{c|c|c} \mathbf{B} & \mathbf{AB} & \mathbf{A}^2\mathbf{B} \end{array}\right] = \left[\begin{array}{cc|cc|cc} 1 & 0 & 0 & 1 & 0 & -1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 2 & 0 & -1 & 0 & 1 \end{array}\right]$$

This matrix has rank $n_r = 2 < 3$, and so $(\mathbf{A}, \mathbf{B})$ is uncontrollable.

A basis for $\mathcal{R}_r$ is obtained by taking the first two linearly independent columns of $\mathcal{C}$ to get

$$\mathbf{Q} = \left[\begin{array}{cc|c} v_1 & v_2 & \mathbf{Q}_1 \end{array}\right] = \left[\begin{array}{cc|c} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 2 & 1 \end{array}\right]$$

where the third column was chosen to make $\mathbf{Q}$ nonsingular. For this $\mathbf{Q}$ we get

$$\begin{aligned} \widehat{\mathbf{A}} &= \mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} \\ &= \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 1 & -2 & 1 \end{bmatrix} \begin{bmatrix} 0 & -1 & 1 \\ 1 & -2 & 1 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 2 & 1 \end{bmatrix} \\ &= \left[\begin{array}{cc|c} 0 & 1 & 1 \\ 0 & -1 & 0 \\ \hline 0 & 0 & -2 \end{array}\right] = \left[\begin{array}{c|c} \mathbf{A}_1 & \mathbf{A}_{12} \\ \hline \mathbf{0} & \mathbf{A}_2 \end{array}\right] \end{aligned}$$

and for $\widehat{\mathbf{B}}$ we get

$$\begin{aligned} \widehat{\mathbf{B}} &= \mathbf{Q}^{-1}\mathbf{B} \\ &= \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 1 & -2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} \\ &= \left[\begin{array}{cc} 1 & 0 \\ 0 & 1 \\ \hline 0 & 0 \end{array}\right] = \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{0} \end{bmatrix} \end{aligned}$$

where $(\mathbf{A}_1, \mathbf{B}_1)$ is controllable. The matrix $\mathbf{A}$ has 3 distinct eigenvalues at 0, $-1$, and $-2$. We see that $\mathbf{A}_1$ has eigenvalues 0 and $-1$ which are controllable and the eigenvalue of $\mathbf{A}_2$ is $-2$ which is uncontrollable.

We can also obtain standard forms for unobservable systems. This means we use a similarity transformation on the pair $(\mathbf{A}, \mathbf{C})$ to decouple the system

into observable and unobservable subsystems. We do this by invoking the duality between the controllability matrix $\mathcal{C}$ and the observability matrix $\mathcal{O}$. In particular, define a dual pair $(\mathbf{A}_D, \mathbf{B}_D)$ where $\mathbf{A}_D = \mathbf{A}^T$ and $\mathbf{B}_D = \mathbf{C}^T$ which will not be controllable. We then use the preceding transformation $\mathbf{Q}_D$ to get

$$\begin{aligned}\hat{\mathbf{A}}_D &= \mathbf{Q}_D^{-1}\mathbf{A}_D\mathbf{Q}_D = \begin{bmatrix} \mathbf{A}_{D1} & \mathbf{A}_{D12} \\ 0 & \mathbf{A}_{D2} \end{bmatrix} \\ \hat{\mathbf{B}}_D &= \mathbf{Q}_D^{-1}\mathbf{B}_D = \begin{bmatrix} \mathbf{B}_{D1} \\ 0 \end{bmatrix}\end{aligned}$$

where $(\mathbf{A}_{D1}, \mathbf{B}_{D1})$ is controllable.

Taking the dual again we obtain $(\hat{\mathbf{A}}, \hat{\mathbf{C}})$

$$\begin{aligned}\hat{\mathbf{A}} = \hat{\mathbf{A}}_D^T &= \mathbf{Q}_D^T\mathbf{A}_D^T(\mathbf{Q}_D^T)^{-1} \\ &= \mathbf{Q}_D^T\mathbf{A}(\mathbf{Q}_D^T)^{-1} = \begin{bmatrix} \mathbf{A}_{D1}^T & 0 \\ \mathbf{A}_{D12}^T & \mathbf{A}_{D2}^T \end{bmatrix} \\ \hat{\mathbf{C}} = \hat{\mathbf{B}}_D^T &= \mathbf{B}_D^T(\mathbf{Q}_D^T)^{-1} \\ &= \mathbf{C}(\mathbf{Q}_D^T)^{-1} = \begin{bmatrix} \mathbf{B}_{D1}^T & 0 \end{bmatrix}\end{aligned}$$

where $(\mathbf{A}_{D1}^T, \mathbf{B}_{D1}^T)$ is observable and the similarity transformation we can use for this is $\mathbf{Q} = (\mathbf{Q}_D^T)^{-1}$.

If we then let $\hat{x} = \mathbf{Q}^{-1}x$ then we get

$$\begin{aligned}\begin{bmatrix} \dot{\hat{x}}_1 \\ \dot{\hat{x}}_2 \end{bmatrix} &= \begin{bmatrix} \mathbf{A}_1 & 0 \\ \mathbf{A}_{21} & \mathbf{A}_2 \end{bmatrix}\begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{bmatrix} u \\ y &= \begin{bmatrix} \mathbf{C}_1 & 0 \end{bmatrix}\begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \end{bmatrix} + \mathbf{D}u\end{aligned}$$

which we call the *standard* form for unobservable systems. The $n_o$ eigenvalues of $\mathbf{A}_1$ and its modes are called observable eigenvalues and modes. The $n - n_o$ eigenvalues of $\mathbf{A}_2$ and modes are unobservable eigenvalues and modes.

The standard form can also be used when $\left[\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right]$ is unobservable and uncontrollable. This is called the *Kalman Decomposition*

$$\begin{aligned} \hat{\mathbf{A}} = \mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} &= \begin{bmatrix} \mathbf{A}_{11} & 0 & \mathbf{A}_{13} & 0 \\ \mathbf{A}_{21} & \mathbf{A}_{22} & \mathbf{A}_{23} & \mathbf{A}_{24} \\ 0 & 0 & \mathbf{A}_{33} & 0 \\ 0 & 0 & \mathbf{A}_{43} & \mathbf{A}_{44} \end{bmatrix} \\ \hat{\mathbf{B}} = \mathbf{Q}^{-1}\mathbf{B} &= \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \\ 0 \\ 0 \end{bmatrix} \\ \hat{\mathbf{C}} = \mathbf{C}\mathbf{Q} &= \begin{bmatrix} \mathbf{C}_1 & 0 & \mathbf{C}_3 & 0 \end{bmatrix} \end{aligned}$$

where $\left(\begin{bmatrix} \mathbf{A}_{11} & 0 \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}, \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{bmatrix}\right)$ is controllable, $\left(\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{13} \\ 0 & \mathbf{A}_{33} \end{bmatrix}, \begin{bmatrix} \mathbf{C}_1 & \mathbf{C}_3 \end{bmatrix}\right)$ is observable and $(\mathbf{A}_{11}, \mathbf{B}_1, \mathbf{C}_1)$ is controllable and observable.

## 6. Eigenvalue/vector Tests for Controllability/Observability

This section derives eigenvalue-vector tests that are used to identify a system's uncontrollable or unobservable modes. These methods can be simpler to use than converting the system to its standard form.

Let us assume there exists a row vector $\widehat{v} \neq 0$ and $\lambda \in \mathbb{C}$ such that

$$\widehat{v}\left[\begin{array}{c|c} \lambda\mathbf{I} - \mathbf{A} & \mathbf{B} \end{array}\right] = 0 \tag{40}$$

This would mean that $\widehat{v}\mathbf{A} = \lambda\widehat{v}$, so that $(\lambda, \widehat{v})$ is a left eigenvalue/vector pair for $\mathbf{A}$. It also means that $\widehat{v}\mathbf{B} = 0$. We can use both observations to see that

$$\widehat{v}\mathbf{A}\mathbf{B} = \lambda\widehat{v}\mathbf{B} = 0$$

and we can then use mathematical induction to see that

$$\widehat{v}\mathbf{A}^k\mathbf{B} = 0$$

for all $k \geq 0$. This means, therefore that

$$\widehat{v}\mathcal{C} = \widehat{v}\begin{bmatrix} \mathbf{A} & \mathbf{AB} & \cdots & \mathbf{A}^{n-1}\mathbf{B} \end{bmatrix} = 0$$

Since $\widehat{v} \neq 0$, we can conclude the system is not completely controllable. So we have just shown that having such a $\widehat{v}$ that satisfies condition (40) if sufficient for uncontrollability.

We can also establish the necessity of this condition (40). In particular, assume that $(\mathbf{A}, \mathbf{B})$ is uncontrollable. We can assume wlog that the realization is in its standard form with

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{A}_{12} \\ \mathbf{0} & \mathbf{A}_2 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{0} \end{bmatrix}$$

with $(\mathbf{A}_1, \mathbf{B}_1)$ being a controllable pair. Now let $\lambda$ be an uncontrollable eigenvalue and let $\widehat{v} = \begin{bmatrix} 0 & \alpha \end{bmatrix}$ where

$$\alpha(\lambda\mathbf{I} - \mathbf{A}_2) = 0$$

This means that condition (40) can be written as

$$\widehat{v}\left[\begin{array}{c|c} \lambda\mathbf{I} - \mathbf{A} & \mathbf{B} \end{array}\right] = \left[\begin{array}{c|c|c} 0 & \alpha(\lambda\mathbf{I} - \mathbf{A}_2) & 0 \end{array}\right] = 0$$

thereby showing that condition (40) is also necessary for uncontrollability. Similar arguments can be used to obtain a second eigenvalue/vector test for unobservability. These two results are summarized as the Popov-Belevich-Hautus (PBH) test for uncontrollability and unobservability.

THEOREM 27. *The pair $(\mathbf{A}, \mathbf{B})$ is uncontrollable if and only if there exists a complex-valued row vector $\widehat{v} \neq 0$ such that*

$$\widehat{v}\left[\begin{array}{c|c} \lambda\mathbf{I} - \mathbf{A} & \mathbf{B} \end{array}\right] = 0$$

*for some $\lambda \in \mathbb{C}$.*

*The pair* $(\mathbf{A}, \mathbf{C})$ *is unobservable if and only if there exists a column vector* $v \neq 0$ *such that*

$$\left[\begin{array}{c} \lambda\mathbf{I} - \mathbf{A} \\ \hline \mathbf{C} \end{array}\right] v = 0$$

*where* $\lambda \in \mathbb{C}$.

**Example:** Consider

$$\dot{x} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} x + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u$$

The $\mathbf{A}$ matrix has an eigenvalue at $1$ with multiplicity $2$. Note that $(\mathbf{A}, \mathbf{B})$ is already in standard form for uncontrollable systems and we can clearly see that one of the repeated eigenvalues at $1$ is uncontrollable since

$$\left[\begin{array}{c|c} \lambda\mathbf{I} - \mathbf{A} & \mathbf{B} \end{array}\right]_{\lambda=1} = \left[\begin{array}{cc|c} 0 & -1 & 1 \\ 0 & 0 & 0 \end{array}\right]$$

has a nonzero left eigenvector $\hat{v} = \begin{bmatrix} 0 & 1 \end{bmatrix}$. So by the PBH test we know this eigenvalue is uncontrollable.

But, we also know the other eigenvalue at $1$ is controllable because this system is in its standard form. This example shows that the PBH eigenvalue test can only detect if one of the repeated eigenvalues is uncontrollable. It cannot be used to detect if the other eigenvalue is controllable.

The proof for the PBH condition suggests that it should be possible to test for controllability/observability directly from the eigenvalues of $\mathbf{A}$. This observation is summarized in the following theorem

THEOREM 28. *The pair* $(\mathbf{A}, \mathbf{B})$ *is controllable if and only if*

$$\operatorname{rank}\left[\begin{array}{c|c} \lambda\mathbf{I} - \mathbf{A} & \mathbf{B} \end{array}\right] = n$$

*for all* $\lambda \in \mathbb{C}$. *If* $\lambda_i$ *is an uncontrollable eigenvalue value of* $\mathbf{A}$, *then* $\operatorname{rank}\left[\begin{array}{c|c} \lambda_i\mathbf{I} - \mathbf{A} & \mathbf{B} \end{array}\right] < n$.

*The pair* $(\mathbf{A}, \mathbf{C})$ *is observable if and only if*

$$\operatorname{rank}\left[\begin{array}{c} \lambda\mathbf{I}-\mathbf{A} \\ \hline \mathbf{C} \end{array}\right] = n$$

*for all* $\lambda \in \mathbb{C}$*. If* $\lambda_i$ *is an unobservable eigenvalue of* $\mathbf{A}$*, then* $\operatorname{rank}\left[\begin{array}{c} \lambda_i\mathbf{I}-\mathbf{A} \\ \hline \mathbf{C} \end{array}\right] < n$*.*

**Example:** Consider the preceding example of the system

$$\begin{aligned} \dot{x} &= \begin{bmatrix} 0 & -1 & 1 \\ 1 & -2 & 1 \\ 0 & 1 & -1 \end{bmatrix} x + \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} u \\ y &= \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} x \end{aligned}$$

Form the matrix

$$\left[\begin{array}{c|c} \lambda\mathbf{I}-\mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & 0 \end{array}\right] = \left[\begin{array}{ccc|cc} \lambda & 1 & -1 & 1 & 0 \\ -1 & \lambda+2 & -1 & 1 & 1 \\ 0 & -1 & \lambda+1 & 1 & 2 \\ \hline 0 & 1 & 0 & 0 & 0 \end{array}\right]$$

The eigenvalues of $\mathbf{A}$ are $0$, $-1$, and $-2$. We can readily see that the only ways the PBH matrices lose rank is when

$$\begin{aligned} \operatorname{rank}\left[\begin{array}{c} \lambda\mathbf{I}-\mathbf{A} \\ \hline \mathbf{C} \end{array}\right]_{\lambda=-1} &= \operatorname{rank}\left[\begin{array}{ccc} -1 & 1 & -1 \\ -1 & 1 & -1 \\ 0 & -1 & 0 \\ \hline 0 & 1 & 0 \end{array}\right] = 2 \\ \operatorname{rank}\left[\begin{array}{c|c} \lambda\mathbf{I}-\mathbf{A} & \mathbf{B} \end{array}\right]_{\lambda=-2} &= \operatorname{rank}\left[\begin{array}{ccc|cc} -2 & 1 & -1 & 1 & 0 \\ -1 & 0 & -1 & 1 & 1 \\ 0 & -1 & -1 & 1 & 2 \end{array}\right] = 2 \end{aligned}$$

which implies the eigenvalue at $-1$ is unobservable and the eigenvalue at $-2$ is uncontrollable.

## 7. Controllable/Observable Realizations

Canonical realizations are state space realizations that have a special useful form. In prior lectures, we introduced two such realizations, those based on companion matrices and those based on diagonal or Jordan matrices. This section takes a closer look at the companion realizations and discusses their relationship to the controllability and observability matrices.

Consider the system

$$\begin{aligned} \dot{x} &= \mathbf{A}x + \mathbf{B}u \\ y &= \mathbf{C}x + \mathbf{D}u \end{aligned}$$

and let $(\mathbf{A}, \mathbf{B})$ be controllable so $\mathrm{rank}(\mathcal{C}) = n$. Assume that

$$\mathrm{rank}(\mathbf{B}) = m \leq n$$

In other words we assume $\mathbf{B}$ has full column rank. We will show how to find the similarity transformation that takes this realization to its controller canonical form.

The controller companion realization is

$$\begin{aligned} \mathbf{A}_c &= \begin{bmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ -\alpha_0 & -\alpha_1 & \cdots & -\alpha_{n-1} \end{bmatrix} \\ \mathbf{B}_c &= \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \\ \mathbf{C}_c &= \text{no particular structure} \end{aligned}$$

where $\alpha_i$ $(i = 0, \ldots, n-1)$ are coefficients of $\mathbf{A}$'s characteristic polynomial

$$\alpha(s) = \det(s\mathbf{I} - \mathbf{A}) = s^n + \alpha_{n-1}s^{n-1} + \cdots + \alpha_1 s + \alpha_0$$

The similarity transformation taking us to this form is

$$\mathbf{P} = \begin{bmatrix} q \\ q\mathbf{A} \\ \vdots \\ q\mathbf{A}^{n-1} \end{bmatrix}$$

where $q$ is the $n$th row (i.e. last row) of the inverse, $\mathcal{C}^{-1}$, of the original system's controllability matrix.

We will verify this assertion by direct computation. Note that

$$\begin{aligned} q\mathbf{A}^{i-1}\mathbf{B} &= 0, \quad \text{for } i = 1, \ldots, n-1 \\ q\mathbf{A}^{n-1}\mathbf{B} &= 1 \end{aligned}$$

This assertion may be verified from the definition of $q$. The above relations can then be seen to imply,

$$q\mathcal{C} = \begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix}.$$

Extending the preceding computation to the product $\mathbf{P}\mathcal{C}$ gives

$$\begin{aligned} \mathbf{P}\mathcal{C} &= \mathbf{P}\begin{bmatrix} \mathbf{B} & \mathbf{AB} & \cdots & \mathbf{A}^{n-1}\mathbf{B} \end{bmatrix} \\ &= \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 1 & x \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 1 & \cdots & x & x \\ 1 & x & \cdots & x & x \end{bmatrix} = \mathcal{C}_c \end{aligned}$$

This shows that

$$\det(\mathbf{P}\mathcal{C}) = \det(\mathbf{P})\det(\mathcal{C}) \neq 0$$

if and only if $\det(\mathcal{C}) \neq 0$ and $\det(\mathbf{P})\neq 0$. The first holds because the original system is controllable and the second condition holds because $\mathbf{P}$ is a similarity transformation (i.e. nonsingular). In view of our expansion for

$\mathbf{P}\mathcal{C}$, we can now use the Cayley-Hamilton theorem to show that

$$\mathbf{A}_c\mathbf{P} = \begin{bmatrix} q\mathbf{A} \\ \vdots \\ q\mathbf{A}^{n-1} \\ q\mathbf{A}^n \end{bmatrix} = \mathbf{P}\mathbf{A}$$

Similar arguments can be used to show

$$\mathbf{P}\mathbf{B} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} = \mathbf{B}_c$$

If we consider an LTI system with multiple inputs, $m > 1$, then the controllability matrix

$$\mathcal{C} = \begin{bmatrix} \mathbf{B} & \mathbf{A}\mathbf{B} & \cdots & \mathbf{A}^{n-1}\mathbf{B} \end{bmatrix}$$

is no longer square. But if $(\mathbf{A}, \mathbf{B})$ is controllable, we can still find $n$ linearly independent columns of $\mathcal{C}$. There are however many ways of choosing these linearly independent columns.

Let $\mathbf{B} = \begin{bmatrix} b_1 & \cdots & b_m \end{bmatrix}$ where $b_i \in \mathbb{R}^n$ is a column of $\mathbf{B}$. We now rearrange the controllability matrix

$$\mathcal{C} = \begin{bmatrix} b_1 & \cdots & b_m & \mathbf{A}b_1 & \cdots & \mathbf{A}b_m & \cdots & \mathbf{A}^{n-1}b_1 & \cdots & \mathbf{A}^{n-1}b_m \end{bmatrix}$$

into the form

$$\overline{\mathcal{C}} = \begin{bmatrix} b_1 & \mathbf{A}b_1 & \cdots & \mathbf{A}^{\mu_1-1}b_1 & \cdots & b_m & \mathbf{A}b_m & \cdots & \mathbf{A}^{\mu_m-1}b_m & \cdots \end{bmatrix}$$

where $\mu_i$ is the number of the first linearly independent columns of the matrix

$$\begin{bmatrix} b_i & \mathbf{A}b_i & \cdots & \mathbf{A}^{n-1}b_i \end{bmatrix}.$$

We call $\mu_i$ the *controllability index* of $b_i$. Note that $\sum_{i=1}^m \mu_i = n$ and the largest controllability index $\mu = \max_i\{\mu_i\}$ is called the controllability index of the system.

We can now form the inverse of $\overline{\mathcal{C}}$ as

$$\overline{\mathcal{C}}^{-1} = \begin{bmatrix} \vdots \\ q_1 \\ \vdots \\ q_m \end{bmatrix}$$

where $q_k$ is the $\sigma_k$th row of $\overline{\mathcal{C}}^{-1}$ and $\sigma_k = \sum_{i=1}^{k} \mu_i$. We can then use these rows to construct a nonsingular matrix that transforms $(\mathbf{A}, \mathbf{B})$ to its MIMO controller canonical form. The similarity transformation is

$$\mathbf{T} = \begin{bmatrix} q_1 \\ q_1\mathbf{A} \\ \vdots \\ q_1\mathbf{A}^{\mu_1 - 1} \\ \vdots \\ q_m \\ q_m\mathbf{A} \\ \vdots \\ q_m\mathbf{A}^{\mu_m - 1} \end{bmatrix}$$

The associated controller canonical form is

$$(\mathbf{A}_c, \mathbf{B}_c) = \left( \left[\begin{array}{cccc|cccc|cccc} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ x & x & x & x & x & x & x & x & x & x & x & x \\ \hline 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ x & x & x & x & x & x & x & x & x & x & x & x \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ x & x & x & x & x & x & x & x & x & x & x & x \end{array}\right], \left[\begin{array}{cccc|cccc} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & x & x & x \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & x & x & x \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & x & x & x \end{array}\right] \right)$$

Note that $\mathbf{C}_c$ has no particular structure.

Consider the system

$$\begin{aligned} \dot{x} &= \mathbf{A}x + \mathbf{B}u \\ y &= \mathbf{C}x + \mathbf{D}u \end{aligned}$$

where $(\mathbf{A}, \mathbf{C})$ is observable with $\mathbf{C}$ being a $p \times n$ matrix with full row rank $p \leq n$.

A similar companion realization called the observer companion form exists when the realization is observable. The observer companion form when $p = 1$ is

$$\begin{aligned} \mathbf{A}_o &= \begin{bmatrix} 0 & \cdots & 0 & -\alpha_0 \\ 1 & \cdots & 0 & -\alpha_1 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 1 & -\alpha_{n-1} \end{bmatrix} \\ \mathbf{C}_o &= \begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix} \end{aligned}$$

where $\alpha_i$ denote the coefficients of the characteristic polynomial

$$\alpha(s) = \det(s\mathbf{I} - \mathbf{A}) = s^n + \alpha_{n-1}s^{n-1} + \cdots + \alpha_1 s + \alpha_0$$

The similarity transformation giving this canonical form is

$$\mathbf{Q} = \begin{bmatrix} \tilde{q} & \mathbf{A}\tilde{q} & \cdots & \mathbf{A}^{n-1}\tilde{q} \end{bmatrix}$$

where $\tilde{q}$ is the $n$th column of the inverse, $\mathcal{O}^{-1}$, of the original system's observability matrix. The derivation of this form is similar to what we did above for the controller companion form.

**Example:** Consider the system

$$\begin{aligned} \dot{x} &= \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{bmatrix} x \\ y &= \begin{bmatrix} 1 & -1 & 1 \end{bmatrix} x \end{aligned}$$

The observability matrix and its inverse are

$$\begin{aligned}\mathcal{O} &= \begin{bmatrix} \mathbf{C} \\ \mathbf{CA} \\ \mathbf{CA}^2 \end{bmatrix} = \begin{bmatrix} 1 & -1 & 1 \\ -1 & -1 & -2 \\ 1 & -1 & 4 \end{bmatrix} \\ \mathcal{O}^{-1} &= \begin{bmatrix} 1 & -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{3} & -\frac{1}{2} & -\frac{1}{6} \\ -\frac{1}{3} & 0 & \frac{1}{3} \end{bmatrix}\end{aligned}$$

So the transformation matrix is

$$\mathbf{Q} = \begin{bmatrix} \tilde{q} & \mathbf{A}\tilde{q} & \mathbf{A}^2\tilde{q} \end{bmatrix} = \begin{bmatrix} -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{6} & -\frac{1}{6} & -\frac{1}{6} \\ \frac{1}{3} & -\frac{2}{3} & \frac{4}{3} \end{bmatrix}$$

and so the observer companion matrices are

$$\begin{aligned}\mathbf{A}_o &= \mathbf{Q}^{-1}\mathbf{AQ} = \begin{bmatrix} 0 & 0 & 2 \\ 1 & 0 & 1 \\ 0 & 1 & -2 \end{bmatrix} \\ \mathbf{C}_o &= \mathbf{CQ} = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}\end{aligned}$$

## 8. Controllability of Modal Realizations

Consider a modal realization of a SISO LTI system with $n$ distinct eigenvalues.

$$\mathbf{G} \overset{s}{=} \left[\begin{array}{c|c} \mathbf{A}_d & \mathbf{B}_d \\ \hline \mathbf{C}_d & 0 \end{array}\right] = \left[\begin{array}{ccc|c} \lambda_1 & \cdots & 0 & b_1 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \lambda_n & b_n \\ \hline c_1 & \cdots & c_n & 0 \end{array}\right]$$

The controllability matrix for this realization is

$$\mathcal{C}_d = \begin{bmatrix} \mathbf{B}_d & \mathbf{A}_d\mathbf{B}_d & \cdots & \mathbf{A}_d^{n-1}\mathbf{B}_d \end{bmatrix}$$

Note that

$$\mathbf{A}_d\mathbf{B}_d = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} = \begin{bmatrix} \lambda_1 b_1 \\ \vdots \\ \lambda_n b_n \end{bmatrix}$$

If we continue inductively, it can be shown that

$$\mathbf{A}_d^k\mathbf{B}_d = \begin{bmatrix} \lambda_1^k b_1 \\ \vdots \\ \lambda_n^k b_n \end{bmatrix}$$

for $k \geq 0$, which means that the controllability matrix for a modal realization whose eigenvalues are distinct has the form

$$\begin{aligned} \mathcal{C}_d &= \begin{bmatrix} b_1 & \lambda_1 b_1 & \cdots & \lambda_1^{n-1} b_1 \\ \vdots & \vdots & & \vdots \\ b_n & \lambda_n b_n & \cdots & \lambda_n^{n-1} b_n \end{bmatrix} \\ &= \begin{bmatrix} b_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & b_n \end{bmatrix} \begin{bmatrix} 1 & \lambda_1 & \lambda_1^2 & \cdots & \lambda_1^{n-1} \\ 1 & \lambda_2 & \lambda_2^2 & \cdots & \lambda_2^{n-1} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & \lambda_n & \lambda_n^2 & \cdots & \lambda_n^{n-1} \end{bmatrix} \\ &= \begin{bmatrix} b_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & b_n \end{bmatrix} \mathbf{V} \end{aligned}$$

The matrix

$$\mathbf{V} = \begin{bmatrix} 1 & \lambda_1 & \lambda_1^2 & \cdots & \lambda_1^{n-1} \\ 1 & \lambda_2 & \lambda_2^2 & \cdots & \lambda_2^{n-1} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & \lambda_n & \lambda_n^2 & \cdots & \lambda_n^{n-1} \end{bmatrix}$$

is called a *Vandermonde* matrix and one can show using induction to show that

$$\det(\mathbf{V}) = \prod_{1 \leq i < j \leq n} (\lambda_j - \lambda_i)$$

This means that the determinant of the modal realization's controllability matrix is

$$\begin{aligned}\det(\mathcal{C}_d) &= \det\left(\begin{bmatrix} b_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & b_n \end{bmatrix}\right)\det(\mathbf{V}) \\ &= \left(\prod_{i=1}^{n} b_i\right)\left(\prod_{1\leq i<j\leq n}(\lambda_j - \lambda_i)\right)\end{aligned}$$

This determinant is nonzero if and only if $\lambda_i \neq \lambda_j$ for $i \neq j$ (i.e. no repeated roots) and $b_i \neq 0$ for all $i = 1, 2, \ldots, n$. From the preceding argument we can therefore conclude that

- A modal realization is controllable if and only if $\lambda_i \neq \lambda_j$ for $i \neq j$ and $b_i \neq 0$ for all $i = 1, 2, \ldots, n$.
- In a similar manner one can say the modal realization is observable if and only if $\lambda_i \neq \lambda_j$ for $i \neq j$ and $c_i \neq 0$ for all $i = 1, 2, \ldots, n$.

Let us now consider what impact uncontrollability might have on the transfer function of a systems with a given modal realization. Consider the transfer function of the $n$ dimensional modal realization with distinct eigenvalues

$$\mathbf{G}(s) = \mathbf{C}_d(s\mathbf{I} - \mathbf{A}_d)^{-1}\mathbf{B}_d = \sum_{i=1}^{n} \frac{b_i c_i}{s - \lambda_i}$$

Since the system eigenvalues are not repeated, then the realization can only be uncontrollable if $b_i = 0$ for some $i = 1, 2, \ldots, n$. The realization can only be unobservable if $c_i = 0$ for some $i = 1, 2, \ldots, n$. If this is the case (i.e. the realization is uncontrollable or unobservable) then the preceding partial fraction expansion only has $r < n$ terms

$$\mathbf{G}(s) = \sum_{j=1}^{r} \frac{b_{i_j} c_{i_j}}{s - \lambda_{i_j}}$$

where $i_1, \ldots, i_r$ are distinct integers drawn from $\{1, 2, \ldots, n\}$. We could write the transfer function as the ratio of two $n$-th order polynomials

$$\begin{aligned}\mathbf{G}(s) &= \frac{b(s)}{a(s)} = \frac{b_{n-1}s^{n-1} + b_{n-2}s^{n-2} + \cdots + b_1 s + b_0}{s^n + a_{n-1}s^{n-1} + \cdots + a_1 s + a_0} \\ &= \frac{b(s)}{\det(s\mathbf{I} - \mathbf{A}_d)}\end{aligned}$$

But recognizing that there are only $r$ nonzero terms in the partial fraction expansion, we can also see that the transfer function may be written as the ratio of two $r$th order polynomials

$$\mathbf{G}(s) = \frac{\beta(s)}{\alpha(s)} = \frac{\beta_{r-1}s^{r-1} + \beta_{r-2}s^{r-2} + \cdots + \beta_1 s + \beta_0}{s^r + \alpha_{r-1}s^{r-1} + \cdots + \alpha_1 s + \alpha_0}$$

where $r < n$. Because $r$ is strictly less than $n$ we can conclude that the $n$th order polynomials $a(s)$ and $b(s)$ share a common root. Since the zeros of $\mathbf{G}(s)$ are the roots of $b(s) = 0$ and the poles of $\mathbf{G}(s)$ are the roots of $a(s) = 0$, we can conclude there is a *pole-zero cancellation* in the original $n$th order transfer function $b(s)/a(s)$ that gives rise to the $r$th order transfer function $\beta(s)/\alpha(s)$. This means that a modal realization with distinct modes is uncontrollable or unobservable if and only if its transfer function $\mathbf{G}(s) = \frac{b(s)}{a(s)}$ has a pole-zero cancellation.

**Remark:** This pole zero cancellation impacts the input-output stability of the system. Note that the zeros of $\det(s\mathbf{I} - \mathbf{A})$ determine whether the origin of the state space realization is asymptotically stable. However, the input-output stability of the system is determined by the transfer function $\beta(s)/\alpha(s)$ obtained after cancelling out the common factors between the numerator and polynomial. This can be seen by an examination of the partial fraction expansion for the transfer function. This means that if an unstable pole of the realization is canceled by a zero it will not impact the input-output behavior of the system. In other words, it is possible for an LTI system to be input-output stable and yet not Lyapunov stable.

The following theorem considers an $n$th order strictly proper transfer function for a SISO system and notes that if one of its $n$th order realizations

is controllable/observable, then all of its $n$th order realizations are controllable/observable.

THEOREM 29. *If a transfer function*

$$\mathbf{G}(s) = \frac{b(s)}{a(s)} = \frac{b_{n-1}s^{n-1} + \cdots + b_0}{s^n + a_{n-1}s^{n-1} + \cdots + a_0}$$

*has one $n$th order realization that is controllable and observable, then all of its $n$th order realizations are controllable and observable.*

**Proof:** Consider two $n$th order realizations of the transfer function $\mathbf{G}(s)$, that we denote as $\left[\begin{array}{c|c} \mathbf{A}_1 & \mathbf{B}_1 \\ \hline \mathbf{C}_1 & 0 \end{array}\right]$ and $\left[\begin{array}{c|c} \mathbf{A}_2 & \mathbf{B}_2 \\ \hline \mathbf{C}_2 & 0 \end{array}\right]$. Assume that $\left[\begin{array}{c|c} \mathbf{A}_1 & \mathbf{B}_1 \\ \hline \mathbf{C}_1 & 0 \end{array}\right]$ is controllable and observable. One can show that

$$\mathcal{O}(\mathbf{A}_1, \mathbf{C}_1)\mathcal{C}(\mathbf{A}_1, \mathbf{B}_1) = \mathcal{O}(\mathbf{A}_2, \mathbf{C}_2)\mathcal{C}(\mathbf{A}_2, \mathbf{B}_2)$$

We assume $\mathcal{O}(\mathbf{A}_1, \mathbf{C}_1)$ and $\mathcal{C}(\mathbf{A}_1, \mathbf{B}_1)$ are nonsingular, so we can immediately conclude $\mathcal{O}(\mathbf{A}_2, \mathbf{C}_2)$ and $\mathcal{C}(\mathbf{A}_2, \mathbf{B}_2)$ are nonsingular, which means the other realization is also controllable and observable. $\diamondsuit$

In view of the above theorem, we will examine the controllability/observability of a companion realization. In particular, let us consider the controller companion realization $\left[\begin{array}{c|c} \mathbf{A}_c & \mathbf{B}_c \\ \hline \mathbf{C}_c & 0 \end{array}\right]$ where $\mathbf{A}_c$ is a suitable companion matrix whose last row contains the coefficients of the matrix' characteristic polynomial. Let $e_i$ denote the $i$th elementary basis vector (i.e. all components of $e_i$ are zero except the $i$th component which is 1). Note that for $1 \leq i \leq n-1$ that

$$\begin{aligned} e_i^T \mathbf{A}_c &= \begin{bmatrix} 0 & \cdots & 1 & \cdots & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-1} \end{bmatrix} \\ &= \begin{bmatrix} 0 & \cdots & 0 & 1 & \cdots & 0 \end{bmatrix} = e_{i+1}^T \end{aligned}$$

where the one component in $e_i$ moves to the $i+1$st place. For $i = n$, we can see that

$$\begin{aligned} e_i^T \mathbf{A}_c &= \begin{bmatrix} 0 & \cdots & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-1} \end{bmatrix} \\ &= \begin{bmatrix} -a_0 & -a_1 & -a_2 & \cdots & -a_{n-1} \end{bmatrix} \end{aligned}$$

We can therefore conclude that

$$\begin{aligned} e_1^T b(\mathbf{A}_c) &= \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix} \left( b_{n-1}\mathbf{A}_c^{n-1} + \cdots + b_0 \mathbf{I} \right) \\ &= b_{n-1} e_1^T \mathbf{A}_c^{n-1} + \cdots + b_0 e_1^T \\ &= b_{n-1} e_2^T \mathbf{A}_c^{n-2} + \cdots + b_0 e_1^T \\ &= b_{n-1} e_n^T + b_{n-2} e_{n-1}^T + \cdots + b_0 e_1^T \\ &= \begin{bmatrix} b_0 & b_1 & \cdots & b_{n-1} \end{bmatrix} = \mathbf{C}_c \end{aligned}$$

In a similar way we can show that

$$\begin{aligned} e_2^T b(\mathbf{A}_c) &= e_1^T \mathbf{A}_c b(\mathbf{A}_c) \\ &= e_1^T b(\mathbf{A}_c) \mathbf{A}_c = \mathbf{C}_c \mathbf{A}_c \end{aligned}$$

and continuing inductively for $e_i^T b(\mathbf{A}_c)$ with $i = 3, \ldots, n-1$ we can show the observability matrix of the controller companion realization is

$$\mathcal{O}_c = \begin{bmatrix} e_1^T \\ \vdots \\ e_n^T \end{bmatrix} b(\mathbf{A}_c) = \mathbf{I} b(\mathbf{A}_c) = b(\mathbf{A}_c)$$

This means that $\mathcal{O}_c$ is nonsingular if and only if $\det(b(\mathbf{A}_c)) \neq 0$. Or rather that the controller companion realization of a system is observable if and only if $\det(b(\mathbf{A}_c)) \neq 0$.

Note however, that

$$\det(b(\mathbf{A}_c)) = \prod_{i=1}^{n} b(\lambda_i)$$

where $\lambda_i$ is the $i$th eigenvalue of $\mathbf{A}_c$. By construction, we also know that

$$a(\lambda_i) = \det(\lambda_i \mathbf{I} - \mathbf{A}_c) = 0$$

So we can say that $b(\lambda_i) = 0$ for some $i$ in $1, 2, \ldots, n$ if and only if $\lambda_i$ is a root of both the $b(s)$ and the characteristic polynomial $\det(s\mathbf{I} - \mathbf{A}_c) = a(s)$. In other words, $\det(b(\mathbf{A}_c)) = 0$ if and only if $a(s)$ and $b(s)$ have a common root. (pole zero cancellation).

The preceding observations motivative the following conventions. Consider two polynomials with distinct roots that we can factor as

$$\begin{aligned} a(s) &= (s - \lambda_1)(s - \lambda_2) \cdots (s - \lambda_n) \\ b(s) &= (s - \mu_1)(s - \mu_2) \cdots (s - \mu_r) \end{aligned}$$

We say the two polynomials are *coprime* if and only if the largest common factor between them is 1. Clearly if $a(s)$ and $b(s)$ are coprime then they do not have a common zero and we can therefore conclude

THEOREM 30. *The $n$th order controller* companion *realization of a strictly proper SISO transfer function $\mathbf{G}(s) = \frac{b(s)}{a(s)}$ is observable if and only if $a(s)$ and $b(s)$ are coprime.*

We say a transfer function $\mathbf{G}(s) = \frac{b(s)}{a(s)}$ is *irreducible* if and only if $a(s)$ and $b(s)$ are coprime. This leads to the following theorem.

THEOREM 31. *A strictly proper transfer function $\mathbf{G}(s) = \frac{b(s)}{a(s)}$ is irreducible if and only if all $n$th order realizations are controllable and observable.*

Finally, we say that a realization of $\mathbf{G}(s)$ is *minimal* if and only if this realization has the smallest state space dimension over all realizations of the transfer function. This definition leads to the following theorem.

THEOREM 32. *A realization of a strictly proper transfer function $\mathbf{G}(s)$ is minimal if and only if $a(s) = \det(s\mathbf{I} - \mathbf{A})$ and $b(s) = \mathbf{C}\left[\mathrm{adj}(s\mathbf{I} - \mathbf{A})\right]\mathbf{B}$ are coprime.*

**Proof:** So we know

$$\mathbf{G}(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} = \frac{\mathbf{C}\left[\text{adj}(s\mathbf{I} - \mathbf{A})\right]\mathbf{B}}{\det(s\mathbf{I} - \mathbf{A})} = \frac{b(s)}{a(s)}$$

Suppose $\left[\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & 0 \end{array}\right]$ is minimal but $\frac{b(s)}{a(s)}$ is not irreducible. This implies there is a lower order transfer function obtained by cancelling the common factors. That lower order transfer function would have a realization whose dimensional is less than that of $\left[\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & 0 \end{array}\right]$ which would contradict the minimality assumption.

Conversely assume that $\left[\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & 0 \end{array}\right]$ is not minimal but that $\frac{b(s)}{a(s)}$ is irreducible. So there is a realization of lower dimensionality. But the transfer function of that lower dimensional realization has an order less than that of $b(s)/a(s)$ hence contradicting the assumption that $b(s)/a(s)$ was irreducible. $\diamondsuit$

We can now conclude with the following theorem, which follows directly from our preceding theorems.

THEOREM 33. $\left[\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & 0 \end{array}\right]$ *is minimal if and only if* $(\mathbf{A}, \mathbf{B})$ *is controllable and* $(\mathbf{A}, \mathbf{C})$ *is observable.*

**Remark:** The preceding discussion focused on minimality of realizations for SISO systems with distinct eigenvalues. We focused on this because the derivation of the results is more easily seen. These results also extend to MIMO systems with repeated eigenvalues, but the proof is more involved and can be found in Antsaklis and Michel (2006).

**Example:** Consider the state space realization

$$\mathbf{G} \stackrel{s}{=} \left[\begin{array}{cccc|c} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 2 & 0 & -2 & 1 \\ \hline 1 & -1 & 0 & 0 & 0 \end{array}\right]$$

Is the state space realization controllable and/or observable?

Rather than trying to solve this by brute force computation of the controllability and observability matrices. Let us first note that this realization is already in controller companion form. So we immediately know that it is controllable and from this form we can immediately write down its transfer function

$$\mathbf{G}(s) = \frac{b(s)}{a(s)} = \frac{-(s-1)}{s^4 + 2s^3 - 2s - 1}$$

We divide out $s-1$ to see if there is a common factor

$$\begin{array}{rrrrrr} & & s^3 & +3s^2 & +3s & +1 \\ \hline s-1) & s^4 & +2s^3 & & -2s & -1 \\ & -s^4 & +s^3 & & & \\ \hline & & 3s^3 & & -2s & -1 \\ & & -3s^3 & +3s^2 & & \\ \hline & & & 3s^2 & -2s & -1 \\ & & & -3s^2 & +3s & \\ \hline & & & & s & -1 \\ & & & & -s & +1 \end{array}$$

This shows that $s-1$ is a common factor between $b(s)$ and $a(s)$ since the remainder of the division is zero. We can therefore conclude that

$$\mathbf{G}(s) = \frac{b(s)}{a(s)} = \frac{-(s-1)}{s^4 + 2s^3 - 2s - 1} = \frac{-1}{(s+1)^3}$$

Since there is a pole zero cancellation, we know the realization is not minimal and so the realization is not observable. Note that this was deduced

without resorting to a direct computation of the controllability and observability matrices.

The relative degree to which modes are "controllable" or "unobservable" depends on which state space realization we choose for the system. It therefore makes sense to consider realizations where a given mode has an equal degree of observability and controllability. Such realizations are said to be *balanced* and provide the basis for obtaining reduced order models (ROM) of linear systems. The following discussion on balanced realizations and their use in model reduction was taken from Green and Limebeer (2012).

A continuous-time state space realization $\mathbf{G} \stackrel{s}{=} \left[\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & 0 \end{array}\right]$ is *balanced* if $\mathbf{A}$ is Hurwitz and

$$\begin{aligned} \mathbf{A}\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\mathbf{A}^T + \mathbf{B}\mathbf{B}^T &= 0 \\ \mathbf{A}^T\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\mathbf{A} + \mathbf{C}^T\mathbf{C} &= 0 \end{aligned}$$

in which

$$\boldsymbol{\Sigma} = \left[\begin{array}{ccc} \sigma_1\mathbf{I}_{r_1} & & \\ & \ddots & \\ & & \sigma_m\mathbf{I}_{r_m} \end{array}\right]$$

with $\sigma_i \neq \sigma_j$ when $i \neq j$ and $\sigma_i > 0$ for all $i = 1, 2, \ldots, m$. Note that $n = r_1 + \cdots + r_m$ and $r_i$ is the multiplicity of $\sigma_i$. We say the realization is an *ordered* balanced realization if $\sigma_1 > \sigma_2 > \cdots > \sigma_m > 0$. In a balanced realization, the basis for the state space is such that each basis vector is "equally" controllable and observable with the degree of controllability/observability being given by the diagonal entry of $\boldsymbol{\Sigma}$. If $\left[\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & 0 \end{array}\right]$ is a balanced realization and the system's initial state $x_0$ is partitioned as

$x_0 = \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix}$ where $x_i$ is an $r_i \times 1$ vector, one can show that

$$\max_{u \in \mathcal{L}_2} \frac{\int_0^\infty |y(\tau)|^2 d\tau}{\int_{-\infty}^0 |u(\tau)|^2 d\tau} = \sum_{i=1}^m \sigma_i^2 x_i^T x_i$$

In other words, $\sigma_i^2$, may be seen as a measure of the extent to which the corresponding $r_i$ dimensional subspace of the state space transfers energy from past inputs to future outputs. The $\sigma_i$'s are called *Hankel singular values* of the system. The next theorem establishes the existence and uniqueness of balanced realizations.

THEOREM 34. *A given realization* $\left[\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & 0 \end{array}\right]$ *can be transformed to a balanced realization if and only if it is asymptotically stable and minimal. Furthermore, the balanced realization is unique up to an ordering of the* $\sigma_i$*'s and an orthognal matrix* $\mathbf{S}$ *satisfying* $\mathbf{S\Sigma} = \mathbf{\Sigma S}$*. When the realization* $\left[\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & 0 \end{array}\right]$ *is asymptotically stable (i.e.* $\mathbf{A}$ *is Hurwitz) and minimal, then the realization* $\left[\begin{array}{c|c} \mathbf{TAT}^{-1} & \mathbf{TB} \\ \hline \mathbf{CT}^{-1} & 0 \end{array}\right]$ *is balanced if we choose the similarity transformation to be* $\mathbf{T} = \mathbf{\Sigma}^{1/2}\mathbf{U}^T\mathbf{R}^{-1}$*, where* $\mathbf{P} = \mathbf{RR}^T$ *is a Cholesky factorization of* $\mathbf{P}$ *and* $\mathbf{R}^T\mathbf{QR} = \mathbf{U\Sigma}^2\mathbf{U}^T$ *is a singular value decomposition of* $\mathbf{R}^T\mathbf{QR}$ *in which* $\mathbf{P}$ *and* $\mathbf{Q}$ *are the controllability and observability gramians that satisfy the Lyapunov equations*

$$\begin{aligned} \mathbf{AP} + \mathbf{PA}^T + \mathbf{BB}^T &= 0 \\ \mathbf{A}^T\mathbf{Q} + \mathbf{QA} + \mathbf{C}^T\mathbf{C} &= 0 \end{aligned}$$

**Proof:** I'll only approve the existence of the balanced realization. Note that if $\mathbf{P}$ and $\mathbf{Q}$ satisfy the Lyapunov equations in the theorem, then for any nonsingular $\mathbf{T}$ we have

$$\begin{aligned} 0 &= (\mathbf{TAT}^{-1})(\mathbf{TPT}^T) + (\mathbf{TPT}^T)(\mathbf{TAT}^{-1})^T + (\mathbf{TB})(\mathbf{TB})^T \\ 0 &= ((\mathbf{T}^T)^{-1}\mathbf{QT}^{-1})(\mathbf{TAT}^{-1}) + (\mathbf{TAT}^{-1})^T((\mathbf{T}^T)^{-1}\mathbf{QT}^{-1}) + (\mathbf{CT}^{-1})^T(\mathbf{CT}^{-1}) \end{aligned}$$

If the original realization is balanced, it is asymptotically stable by assumption and $\mathbf{\Sigma} > 0$ implies minimality. If the original realization is asymptotically stable and minimal, then it has positive definite controllability and observability gramians, $\mathbf{P}$ and $\mathbf{Q}$, satisfying the two Lyapunov equations. Setting $\mathbf{T} = \mathbf{\Sigma}^{1/2}\mathbf{U}^T\mathbf{R}^{-1}$ gives

$$\begin{aligned} \mathbf{TPT}^T &= (\mathbf{\Sigma}^{1/2}\mathbf{U}^T\mathbf{R}^{-1})\mathbf{R}\mathbf{R}^T((\mathbf{R}^T)^{-1}\mathbf{U}\mathbf{\Sigma}^{1/2}) = \mathbf{\Sigma} \\ (\mathbf{T}^T)^{-1}\mathbf{QT}^{-1} &= (\mathbf{\Sigma}^{-1/2}\mathbf{U}^T\mathbf{R}^T)\mathbf{Q}(\mathbf{RU}\mathbf{\Sigma}^{-1/2}) = \mathbf{\Sigma} \end{aligned}$$

$\diamondsuit$

## 9. Model Reduction

We already noted that modal realizations are useful in the sense that their eigenvalues are less sensitive than the companion canonical forms with respect to perturbations of the system matrices. Another useful aspect of modal forms is their use in *model reduction* [Green and Limebeer (2012)]. Many applications give rise to extremely high order state space realizations. Not all of these modes, however, are of equal importance to the application. So given a state space realization with large dimension $n$, $\mathbf{G} \stackrel{s}{=} \left[\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & 0 \end{array}\right]$, there is value in generating realizations $\hat{\mathbf{G}} \stackrel{s}{=} \left[\begin{array}{c|c} \hat{\mathbf{A}} & \hat{\mathbf{B}} \\ \hline \hat{\mathbf{C}} & 0 \end{array}\right]$ whose dimension $r < n$ and yet the realizations input-output behavior (i.e. transfer function) is very similar to that of the high dimensional system.

One of the easiest ways of generating such reduced order systems is to take a given realization and discard those states that are felt to have little impact on the overall system's behavior. In particular, let us assume the original system ($n$-dimensional) is

$$\begin{aligned} \dot{x} &= \mathbf{A}x + \mathbf{B}u \\ y &= \mathbf{C}x + \mathbf{D}u \end{aligned}$$

Let us assume this is in its modal form with distinct eigenvalues. We take the state $x$ and partition it as

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

where $x_1$ is $r$-dimensional and $x_2$ is $n - r$ dimensional and represents the states of the system we want to discard in forming the reduced model. Because we are discarding states, we also refer to this as *model truncation*. Let us conformally partition the original system matrices with respect to our decomposition of $x$ as

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} \mathbf{C}_1 & \mathbf{C}_2 \end{bmatrix}$$

The truncated system obtained by discarding $x_2$ is then

$$\begin{aligned} \dot{x}_1 &= \mathbf{A}_{11}x_1 + \mathbf{B}_1 u \\ y &= \mathbf{C}_1 x_1 \end{aligned}$$

In particular, if we started with a modal realization with distinct eigenvalues then it should be apparent that

$$\mathbf{A}_{11} = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_r \end{bmatrix}, \quad \mathbf{B}_1 = \begin{bmatrix} \overline{\mathbf{B}}_1 \\ \vdots \\ \overline{\mathbf{B}}_r \end{bmatrix}, \quad \mathbf{C}_1 = \begin{bmatrix} \overline{\mathbf{C}}_1 & \cdots & \overline{\mathbf{C}}_r \end{bmatrix}$$

We are interested in selecting those modes to truncate such that the difference between $\mathbf{G}$ and $\hat{\mathbf{G}}$ is small. How should this be done?

In particular, note that if we use the same input to drive both $\mathbf{G}$ and $\hat{\mathbf{G}}$, we would want their corresponding outputs to be "close" to each other with respect to some appropriate norm. We usually choose the $\mathcal{L}_2$ norm (i.e. energy). Recall that

$$\left\| \mathbf{G} - \hat{\mathbf{G}} \right\|_{\mathcal{H}_\infty} = \sup_{u \in \mathcal{L}_2} \frac{\|\mathbf{G}[u] - \hat{\mathbf{G}}[u]\|_{\mathcal{L}_2}}{\|u\|_{\mathcal{L}_2}}$$

So we want truncate those modes whose transfer of input signal energy to the output is "small". That "energy" transfer is measured by the $\mathcal{H}_\infty$ norm of the error system $\mathbf{G} - \hat{\mathbf{G}}$.

So what is the $\mathcal{H}_\infty$ norm of the error system. When we have a modal realization, we can bound this rather easily. In particular, it should be clear that the error system for a modal realization is

$$\mathbf{G}(s) - \hat{\mathbf{G}}(s) = \sum_{i=r+1}^{n} \frac{\mathbf{C}_i \mathbf{B}_i}{s - \lambda_i}$$

We can then use standard bounding arguments to show that

$$\|\mathbf{G} - \hat{\mathbf{G}}\|_{\mathcal{H}_\infty} \leq \sum_{i=r+1}^{n} \frac{\|\mathbf{C}_i \mathbf{B}_i\|}{\mathrm{Real}(\lambda_i)}$$

In other words, we want to select those $n - r$ modes to discard such that the preceding sum is minimized.

Note that common engineering practice is to discard high frequency modes. In light of the preceding equation, this makes sense if all $\|\mathbf{C}_i \mathbf{B}_i\|$ are the same. In real life, however, the terms $\|\mathbf{C}_i \mathbf{B}_i\|$ will not be the same and they are, in fact, dependent on how we decided to factor the residue term $\mathbf{R}_i$. This suggests that the common engineering practice of discarding high frequency modes (i.e. modal truncation) is not necessarily the best strategy. In particular, we will show that a better strategy is to truncate modes of a *balanced realization* of the system. The key feature of such balanced realizations is that the observability and controllability gramians of the system are the same. In particular, one may say a balanced realization is a particular type of modal realization in which each mode has the same "degree" of observability and controllability.

Model reduction by *balanced truncation* simply applies the truncation operation to a balanced realization $\mathbf{G} \stackrel{s}{=} \left[\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & 0 \end{array}\right]$. In particular, if the

realization is balanced then we can partition $\mathbf{\Sigma}$ as

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_1 & 0 \\ 0 & \mathbf{\Sigma}_2 \end{bmatrix}$$

with

$$\mathbf{\Sigma}_1 = \begin{bmatrix} \sigma_1 \mathbf{I}_{r_1} & & \\ & \ddots & \\ & & \sigma_\ell \mathbf{I}_{r_\ell} \end{bmatrix}, \quad \mathbf{\Sigma}_2 = \begin{bmatrix} \sigma_{\ell+1} \mathbf{I}_{r_{\ell+1}} & & \\ & \ddots & \\ & & \sigma_m \mathbf{I}_{r_m} \end{bmatrix}$$

Note that we don't split states corresponding to a $\sigma_i$ with multiplicity greater than one. If we then partition $\left[\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & 0 \end{array}\right]$ conformally with $\mathbf{\Sigma}$, we obtain a reduced order system $\hat{\mathbf{G}}$ with realization $\left[\begin{array}{c|c} \mathbf{A}_{11} & \mathbf{B}_1 \\ \hline \mathbf{C}_1 & 0 \end{array}\right]$ which is a balanced truncation of $\mathbf{G}$. The following theorems regarding the balanced truncation are stated below without proof.

THEOREM 35. *If* $\left[\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & 0 \end{array}\right]$ *is a balanced realization, then a balanced truncation with realization* $\left[\begin{array}{c|c} \mathbf{A}_{11} & \mathbf{B}_1 \\ \hline \mathbf{C}_1 & 0 \end{array}\right]$ *is also a balanced realization.*

THEOREM 36. [1] *Let* $\mathbf{G}(s) = \mathbf{D} + \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}$ *where* $\left[\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right]$ *is a balanced realization with a balanced truncation* $\left[\begin{array}{c|c} \mathbf{A}_{11} & \mathbf{B}_1 \\ \hline \mathbf{C}_1 & \mathbf{D} \end{array}\right]$ *with* $r$ *modes in which* $r = r_1 + \cdots + r_\ell$. *Then*

$$\|\mathbf{G} - \hat{\mathbf{G}}\|_{\mathcal{H}_\infty} \leq 2(\sigma_{\ell+1} + \cdots + \sigma_m)$$

**Example:** Consider the following system for a flexible structure

$$\mathbf{G}(s) = \sum_{i=1}^{4} k_i \frac{\omega_i^2}{s^2 + 2\zeta_i \omega_i s + \omega_i^2}$$

[1] The proof of this theorem relies on facts about the Hankel singular value that would take some time to develop. For details the reader can refer to (Green and Limebeer, 2012).

| $i$ | $\omega_i$ | $\zeta_i$ | $k_i$ |
|---|---|---|---|
| 1 | 0.5680 | 0.0010 | 0.0165 |
| 2 | 3.9400 | 0.0010 | 0.0020 |
| 3 | 10.5800 | 0.0010 | 0.0100 |
| 4 | 16.1900 | 0.0100 | 0.0002 |

which can be seen as having four vibrational models $\omega_i$ for $i = 1, 2, 3, 4$ where

- Use Matlab to compute the modal canonical form of this system, $\mathbf{G}$, and determine the modal truncation $\hat{\mathbf{G}}_m$ that truncates the two vibrational modes with the highest natural frequencies. Compute $\|\mathbf{G} - \hat{\mathbf{G}}_m\|_{\mathcal{H}_\infty}$.
- Determine the balanced realization of $\mathbf{G}$. Verify that the observability and controllability gramians for the balanced realization are the same. Determine the modal truncation $\hat{\mathbf{G}}$ where the two vibrational modes with smallest Hankel singular values are truncated. Compute $\|\mathbf{G} - \hat{\mathbf{G}}_b\|$ and compare to the truncation error computed in the first part.
- Use Matlab to draw the gain-magnitude plots of $\mathbf{G}$, $\hat{\mathbf{G}}_m$ and $\hat{\mathbf{G}}_b$. Describe the difference between the two different truncation methods?

Let $\mathbf{G}_i = \frac{k_i\omega_i^2}{s^2+2\zeta_i\omega_i s+\omega_i^2}$. Using MATLAB `canon` to construct the modal form for $\mathbf{G} = \sum_{i=1}^{4} \mathbf{G}_i$ yields (note that we have rounded the results to the second least significant digit)

$$\mathbf{G} \stackrel{s}{=} \left[\begin{array}{cccccccc|c} -0.016 & 16.19 & 0 & 0 & 0 & 0 & 0 & 0 & 1.30 \\ -16.19 & -0.16 & 0 & 0 & 0 & 0 & 0 & 0 & 0.01 \\ 0 & 0 & -0.01 & 10.58 & 0 & 0 & 0 & 0 & 1.12 \\ 0 & 0 & -10.58 & -0.011 & 0 & 0 & 0 & 0 & -0.09 \\ 0 & 0 & 0 & 0 & -0.00 & 3.94 & 0 & 0 & -0.22 \\ 0 & 0 & 0 & 0 & -3.94 & -0.00 & 0 & 0 & -0.00 \\ 0 & 0 & 0 & 0 & 0 & 0 & -0.00 & 0.57 & 0.07 \\ 0 & 0 & 0 & 0 & 0 & 0 & -0.57 & -0.00 & 0.00 \\ \hline 0.00 & -0.00 & -0.01 & -0.10 & -0.00 & 0.04 & 0.00 & -0.13 & 0 \end{array}\right]$$

The modal truncation where the vibrational modes with the highest natural frequencies are dropped is $\mathbf{G} = \mathbf{G}_1 + \mathbf{G}_2$ and the modal form for this is

$$\mathbf{G} \stackrel{s}{=} \left[\begin{array}{rrrrr} -0.00 & 3.94 & 0 & 0 & -0.29 \\ -3.94 & -0.00 & 0 & 0 & -0.00 \\ 0 & 0 & -0.00 & 0.57 & 0.13 \\ 0 & 0 & -0.57 & -0.00 & -0.01 \\ \hline -0.00 & -0.03 & -0.00 & -0.07 & 0 \end{array}\right]$$

Figure 2 shows the Bode plot for $\mathbf{G}$ and $\hat{\mathbf{G}}$ as well as the error system $\mathbf{E} = \mathbf{G} - \hat{\mathbf{G}}$. We used `norm` command to identify the $\mathcal{H}_\infty$ truncation error as $5$ ($13.98$ dB) at $\omega = 10.58$ rad/sec.

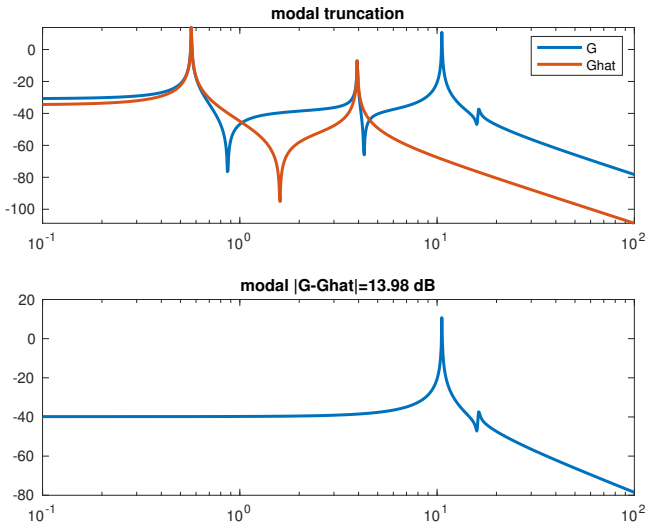



FIGURE 2. Modal Truncation

Now we look at the second part using the balanced realization. We first compute the controllability and observability gramians for the original realization

```
P = lyap(A,B*B');    %AP+PA'+BB'=0
Q = lyap(A',C'*C);   %A'Q+QA+C'C=0
```

and then we compute the transformation

$$\mathbf{T} = \mathbf{\Sigma}^{1/2}\mathbf{U}^T\mathbf{R}^{-1}$$

where $\mathbf{P} = \mathbf{R}\mathbf{R}^{-1}$ and $\mathbf{R}^T Q \mathbf{R} = \mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^T$ is a singular value decomposition. The balanced realization is then

$$(\mathbf{T}\mathbf{A}\mathbf{T}^{-1}, \mathbf{T}\mathbf{B}, \mathbf{C}\mathbf{T}^{-1}\mathbf{D}).$$

I generated the following balanced realization using this method

$$\mathbf{G} \stackrel{s}{=} \left[\begin{array}{rrrrrrrrr} -0.00 & -0.57 & -0.00 & 0.01 & 0.00 & 0.00 & -0.00 & -0.00 & 0.07 \\ 0.57 & -0.00 & -0.01 & 0.00 & 0.00 & 0.00 & -0.00 & -0.0007 & -0.0682 \\ -0.0024 & 0.0097 & -0.0106 & 10.5800 & 0.0048 & 0.0072 & -0.0037 & -0.00 & 0.23 \\ -0.01 & 0.00 & -10.58 & -0.01 & -0.01 & -0.00 & 0.00 & 0.00 & 0.23 \\ 0.00 & -0.00 & 0.00 & 0.01 & -0.00 & -3.94 & 0.01 & 0.0051 & -0.0628 \\ -0.00 & 0.00 & -0.01 & -0.00 & 3.94 & -0.00 & 0.01 & 0.01 & 0.06 \\ -0.00 & 0.00 & -0.00 & -0.00 & 0.01 & -0.01 & -0.16 & -16.1892 & 0.0400 \\ 0.00 & -0.00 & 0.00 & 0.00 & -0.01 & 0.01 & 16.19 & -0.17 & -0.04 \\ \hline 0.07 & 0.07 & 0.23 & -0.23 & -0.06 & -0.06 & 0.04 & 0.04 & 0 \end{array}\right]$$

If we compute the gramians for this realization we essentially obtain the following

$$\begin{aligned} \mathbf{W}_c &= \operatorname{diag}(4.1291, 4.1209, 2.5025, 2.4975, 0.5005, 0.4995, 0.0050, 0.0049) \\ \mathbf{W}_o &= \operatorname{diag}(4.1209, 4.1291, 2.4975, 2.5025, 0.4995, 0.5005, 0.0049, 0.0050) \end{aligned}$$

which is essentially the same (up to reordering) and so we've verified that the realization is balanced.

We now truncate the realization by dropping the last 4 modes in the balanced realization to obtain

$$\hat{\mathbf{G}} \stackrel{s}{=} \left[\begin{array}{rrrrr} -0.00 & -0.57 & -0.00 & 0.01 & 0.07 \\ 0.57 & -0.00 & -0.01 & 0.00 & -0.07 \\ -0.00 & 0.01 & -0.01 & 10.58 & 0.23 \\ -0.01 & 0.00 & -10.58 & -0.01 & 0.23 \\ \hline 0.07 & 0.68 & 0.23 & -0.23 & 0 \end{array}\right]$$

Figure 3 shows the Bode plot for $\mathbf{G}$ and $\hat{\mathbf{G}}$ as well as the error system $\mathbf{E} = \mathbf{G} - \hat{\mathbf{G}}$. We used `norm` command to identify the $\mathcal{H}_\infty$ truncation error as $1$ ($0$ dB) at $\omega = 3.94$ rad/sec. In comparing the modal versus the balanced truncation we see that the balanced truncation produces a lower error. This is accomplished because the balanced truncation retains the two vibrational modes with the largest gain magnitude. The vibration modes with smallest

natural frequency are not the modes with the largest gain magnitudes and so the modal truncation produces a larger model approximation error.

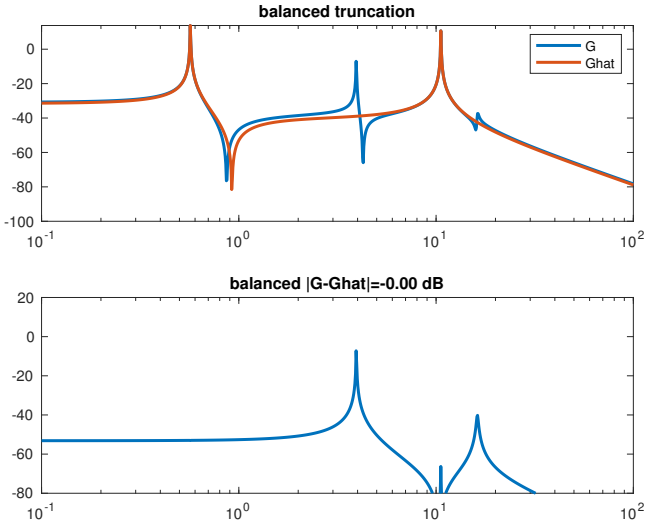



FIGURE 3. Balanced Truncation

CHAPTER 5

# Feedback Theory for Linear Systems

**Feedback** is a useful mechanism that one uses to regulate and stabilize the behavior of a dynamical system. Consider an LTI system

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

Feedback takes the output, $y$, and subtract it from a reference signal, $r$, to form an error signal, $e(t) = y(t) - r(t)$. Taking a linear combination of these error terms generates an input $u(t) = k^T e(t)$, where $k$ is a vector of gains. We are interested in choosing $k$ so the error, $e$, asymptotically goes to zero (stability), thereby *regulating* the output, $y$, to track the reference $r$. This chapter examines feedback mechanisms used to stabilize the states and regulate the outputs of a linear system. These methods are also relevant to the design of state estimators that we refer to as *observers*. The chapter confines itself to continuous-time time-invariant linear systems.

## 1. State Feedback

Consider the continuous-time LTI system whose state space realization is

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

We say the realization can have its eigenvalues *arbitrarily assigned by state feedback* if for any $n$th order polynomial $\alpha_d(s)$, there is a matrix $\mathbf{F} \in \mathbb{R}^{m \times n}$ such that the eigenvalues of $\mathbf{A}+\mathbf{BF}$ are the roots of the polynomial equation

$\alpha_d(s) = 0$. In other words, there is a *state feedback law* of the form

$$u(t) = \mathbf{F}x(t)$$

such that when this $u$ is applied to our system, we obtain

$$\begin{aligned} \dot{x}(t) &= (\mathbf{A} + \mathbf{BF})x(t) \\ y(t) &= \mathbf{C}x(t) \end{aligned}$$

By requiring the characteristic polynomial of $\mathbf{A} + \mathbf{BF}$ to equal $\alpha_d(s)$ for whatever $n$th order polynomial we select, we are using state feedback to arbitrarily place the eigenvalues of the closed loop system matrix, $\mathbf{A} + \mathbf{BF}$. We will show that the eigenvalues of $(\mathbf{A}, \mathbf{B})$ can be freely assigned in and only if $(\mathbf{A}, \mathbf{B})$ is a controllable pair.

To prove this assertion, let us assume the eigenvalues of $(\mathbf{A}, \mathbf{B})$ can be arbitrarily assigned, but that $(\mathbf{A}, \mathbf{B})$ is uncontrollable. Then we know there exists a nonsingular matrix $\mathbf{P}$ that takes the system to its standard uncontrollable form

$$\mathbf{P}^{-1}\mathbf{AP} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{A}_{12} \\ \mathbf{0} & \mathbf{A}_2 \end{bmatrix}, \quad \mathbf{P}^{-1}\mathbf{B} = \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{0} \end{bmatrix}$$

where $(\mathbf{A}_1, \mathbf{B}_1)$ is a controllable pair.

Now consider a matrix $\mathbf{F} \in \mathbb{R}^{n \times m}$ and apply $\mathbf{P}$ to it as

$$\mathbf{FP} = \begin{bmatrix} \mathbf{F}_1 & \mathbf{F}_2 \end{bmatrix}$$

where the block matrices are conformal with the standard form blocks. If we then look at $\mathbf{A} + \mathbf{BF}$ and apply our similarity transformation to it, we get

$$\begin{aligned} \mathbf{P}^{-1}(\mathbf{A} + \mathbf{BF})\mathbf{P} &= \mathbf{P}^{-1}\mathbf{AP} + \mathbf{P}^{-1}\mathbf{BFP} \\ &= \begin{bmatrix} \mathbf{A}_1 & \mathbf{A}_{12} \\ \mathbf{0} & \mathbf{A}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{F}_1 & \mathbf{F}_2 \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{A} + \mathbf{B}_1\mathbf{F}_1 & \mathbf{A}_{12} + \mathbf{B}_1\mathbf{F}_2 \\ \mathbf{0} & \mathbf{A}_2 \end{bmatrix} \end{aligned}$$

The characteristic polynomial is invariant under similarity transformations so with the above block partition we have

$$\begin{aligned}\det(s\mathbf{I}-(\mathbf{A}+\mathbf{BF})) &= \det(s\mathbf{I}-\mathbf{P}^{-1}(\mathbf{A}+\mathbf{BF})\mathbf{P}) \\ &= \det\begin{bmatrix} s\mathbf{I}-(\mathbf{A}_1+\mathbf{B}_1\mathbf{F}_1) & -(\mathbf{A}_{12}+\mathbf{B}_1\mathbf{F}_2) \\ \mathbf{0} & s\mathbf{I}-\mathbf{A}_2 \end{bmatrix} \\ &= \det(s\mathbf{I}-(\mathbf{A}_1+\mathbf{B}_1\mathbf{F}_1))\det(s\mathbf{I}-\mathbf{A}_2)\end{aligned}$$

This means that the eigenvalues of $\mathbf{A}+\mathbf{BF}$ are either eigenvalues of $\mathbf{A}_2$ and or are eigenvalues $\mathbf{A}_1+\mathbf{B}_1\mathbf{F}_1$. Clearly the uncontrollable eigenvalues of $\mathbf{A}_2$ cannot be freely reassigned by state feedback. This contradicts our earlier assumption that the eigenvalues of $(\mathbf{A},\mathbf{B})$ were freely assignable. The contradiction arose because we required $(\mathbf{A},\mathbf{B})$ to also be uncontrollable. We can, therefore, conclude that if the eigenvalues of freely assignable then the system must be controllable.

Controllability is not only necessary for free assignment, it is also sufficient. This assertion can be verified as follows. Let us assume $(\mathbf{A},\mathbf{B})$ is a controllable pair, then it will have a controller companion form

$$(\mathbf{A}_c,\mathbf{B}_c)=\left(\begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-1} \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}\right)$$

Now consider the state feedback gain matrix

$$\mathbf{F}_c = \begin{bmatrix} f_0 & \cdots & f_{n-1} \end{bmatrix}$$

and form $(\mathbf{A}_c+\mathbf{B}_c\mathbf{F}_c)$ as

$$\begin{aligned}\mathbf{A}_c+\mathbf{B}_c\mathbf{F}_c &= \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-1} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \begin{bmatrix} f_0 & \cdots & f_{n-1} \end{bmatrix} \\ &= \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -(a_0-f_0) & -(a_1-f_1) & -(a_2-f_2) & \cdots & -(a_{n-1}-f_{n-1}) \end{bmatrix}\end{aligned}$$

Note that $(\mathbf{A}_c+\mathbf{B}_c\mathbf{F}_c, \mathbf{B}_c)$ is still in controller companion form so that

$$\det(s\mathbf{I}-(\mathbf{A}_c+\mathbf{B}_c\mathbf{F}_c)) = (a_0-f_0)+(a_1-f_1)s+\cdots+(a_{n-1}-f_{n-1})s^{n-1}+s^n$$

So for any desired $n$th order polynomal, $\alpha_d(s)$, there is a set of gains, $\mathbf{F}_c$, such that $\det(s\mathbf{I}-(\mathbf{A}_c+\mathbf{B}_c\mathbf{F}_c)) = \alpha_d(s)$ and so $(\mathbf{A}_c, \mathbf{B}_c)$ has freely assignable eigenvalues. Since eigenvalues are invariant under similarity transformations, we can conclude $(\mathbf{A}, \mathbf{B})$ will also have freely assignable eigenvalues. The preceding discussion can be summarized in the following theorem

THEOREM 37. *Consider a continuous-time LTI system with the pair* $(\mathbf{A}, \mathbf{B})$*, then the pair has eignvalues that are arbitrarily assignable by state feedback if and only if* $(\mathbf{A}, \mathbf{B})$ *is controllable.*

Note that if $\mathbf{P}$ is the similarity transformation taking $(\mathbf{A}, \mathbf{B})$ to its standard form. Once we've selected gains for the standard form, the gains for the original realization are

$$\mathbf{FP} = \mathbf{F}_c$$

In general we want to select $\mathbf{F}$ so that $\mathbf{A}+\mathbf{BF}$ is Hurwitz. We say the pair $(\mathbf{A}, \mathbf{B})$ is *stabilizable* if only if there exists an $\mathbf{F}$ such that all eigenvalues of $(\mathbf{A}+\mathbf{BF})$ have negative real parts. A necessary and sufficient condition for the existence of such a set of gains is that all uncontrollable modes of the system are asymptotically stable.

**Example:** Consider the system

$$\begin{aligned} \dot{x}(t) &= \begin{bmatrix} 1/3 & 1/3 & -2/3 \\ 1/3 & -2/3 & 1/3 \\ -2/3 & -5/3 & -2/3 \end{bmatrix} x(t) + \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix} u(t) \\ y(t) &= \begin{bmatrix} 1 & 1 & -2 \end{bmatrix} x(t) \end{aligned} \tag{41}$$

and determine a linear state feedback law

$$u(t) = \mathbf{F}x(t) + kr \tag{42}$$

where $\mathbf{F}$ is the state feedback gain matrix and $k$ is a scalar gain so that the irreducible transfer function of the closed loop system formed from equations (41-42) is equal to $\frac{1}{s^2+3s+2}$. The other input, $r$, is an exogenous reference signal that is supplied to the system.

Are the eigenvalues of $(\mathbf{A}, \mathbf{B})$ freely assignable? To answer this we first check to see if the pair is controllable or not. Computing the controllability matrix

$$\mathcal{C} = \begin{bmatrix} \mathbf{B} & \mathbf{AB} & \mathbf{A}^2\mathbf{B} \end{bmatrix} = \begin{bmatrix} 1/3 & 0 & 2/3 \\ 1/3 & 0 & -1/3 \\ 1/3 & -1 & 2/3 \end{bmatrix}$$

The eigenvalues of $\mathcal{C}$ are all nonzero and so $\mathcal{C}$ has full rank and we know there exists a gain that can freely assign the eigenvalues.

Following what was done in the proof of the theorem, let us first convert $(\mathbf{A}, \mathbf{B})$ to its standard form. From the preceding chapter we know the desired transformation matrix to standard form is

$$\mathbf{P}^{-1} = \begin{bmatrix} q \\ q\mathbf{A} \\ q\mathbf{A}^2 \end{bmatrix}$$

where $q$ is a row vector obtained from the last row of $\mathcal{C}^{-1}$. We can readily see that

$$\mathcal{C}^{-1} = \begin{bmatrix} 1/3 & 0 & 2/3 \\ 1/3 & 0 & -1/3 \\ 1/3 & -1 & 2/3 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 2 & 0 \\ 1 & 0 & -1 \\ 1 & -1 & 0 \end{bmatrix}$$

so that

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

and so we get

$$\mathbf{P}^{-1} = \begin{bmatrix} q \\ q\mathbf{A} \\ q\mathbf{A}^2 \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 1 & 1 & 1 \end{bmatrix}$$

and converting $(\mathbf{A}, \mathbf{B})$ to controllable companion form is

$$\begin{aligned} \mathbf{A}_c &= \mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & -1 \end{bmatrix} \\ \mathbf{B}_c &= \mathbf{P}^{-1}\mathbf{B} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \\ \mathbf{C}_c &= \mathbf{C}\mathbf{P} = \begin{bmatrix} 1 & 2 & 0 \end{bmatrix} \end{aligned}$$

with $\mathbf{D}_c = 0$.

So now consider

$$\begin{aligned} \dot{x}_c &= \mathbf{A}_c x_c + \mathbf{B}_c(\mathbf{F}_c x_c + kr) \\ &= (\mathbf{A} + \mathbf{B}_c\mathbf{F}_c)x_c + \mathbf{B}_c kr \\ &= \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 + f_{0c} & f_{1c} & -1 + f_{2c} \end{bmatrix} x_c + \begin{bmatrix} 0 \\ 0 \\ k \end{bmatrix} \\ y(t) &= \begin{bmatrix} 1 & 2 & 0 \end{bmatrix} x_c(t) \end{aligned}$$

where $x_c$ is the state of the controller companion realization.

Let us write out the transfer function for this system

$$\mathbf{T}_{ry}(s) = \frac{(2s+1)k_c}{s^3 + (1-f_{2c})s^2 + (-f_{1c})s + (-1-f_{0c})} = \frac{n(s)}{d(s)}$$

We will need to introduce a pole zero cancellation at $-1/2$ so the desired characteristic polynomial we wish to match is

$$\alpha_d(s) = (s+1/2)(s^2+3s+2) = s^3 + 3.5s^2 + 3.5s + 1$$

which suggests we need

$$f_{2c} = -2.5, \quad f_{1c} = -3.5, \quad f_{0c} = -2$$

and so $\mathbf{F}_c = \begin{bmatrix} -2 & -3.5 & -2.5 \end{bmatrix}$. We will also need to select $k = 1/2$ to get the desired numerator.

The actual gains we are going to use in our original system are

$$\begin{aligned} \mathbf{F} &= \mathbf{F}_c\mathbf{P} \\ &= \begin{bmatrix} -2 & -3.5 & -2.5 \end{bmatrix} \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} -4.5 & -4 & 1 \end{bmatrix} \end{aligned}$$

To check our work, compute the transfer function for the state space realization of the closed-loop system

$$T_{ry} \stackrel{s}{=} \left[\begin{array}{c|c} \mathbf{A}+\mathbf{BF} & \mathbf{B}k \\ \hline \mathbf{C} & 0 \end{array}\right] = \left[\begin{array}{ccc|c} -1\frac{1}{6} & -1 & -\frac{1}{3} & \frac{1}{6} \\ -1\frac{1}{6} & -2 & \frac{2}{3} & \frac{1}{6} \\ -2\frac{1}{6} & -3 & -\frac{1}{3} & \frac{1}{6} \\ \hline 1 & 1 & -2 & 0 \end{array}\right]$$

If we compute the transfer functions for this system we see it is

$$T_{ry}(s) = \frac{s+0.5}{(s+2)(s+1)(s+0.5)} = \frac{1}{(s+2)(s+1)}$$

which matches our specification. As expected, the pole we added at $-1/2$ indeed cancelled out the transmission zero at $-1/2$.

The preceding discussion showed how to select the gains $\mathbf{F}$ to arbitrarily reassign the eigenvalues of the closed-loop system. This approach required that we first transform the system to its controllable companion form. A more direct "formula" for pole placement is known as *Ackerman's Formula*. This formula directly computes the state feedback gains required to arbitrarily assign the eigenvalues of $\mathbf{A} + \mathbf{BF}$. The following theorem gives this formula

THEOREM 38. *Assume that* $(\mathbf{A}, \mathbf{B})$ *is an* $n$*-dimensional controllable pair and let* $\alpha_d(s)$ *be an* $n^{\text{th}}$ *order monic polynomial. Then the eigenvalues of* $(\mathbf{A} + \mathbf{BF})$ *will be equal to the roots of* $\alpha_d(s) = 0$ *if and only if*

$$\mathbf{F} = -e_n^T \mathcal{C}^{-1} \alpha_d(\mathbf{A})$$

*where*

$$\begin{aligned} e_n &= \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \in \mathbb{R}^n \\ \alpha_d(s) &= \textit{desired characteristic polynomial} \\ &= s^n + d_{n-1}s^{n-1} + \cdots + d_1 s + d_0 \end{aligned}$$

**Proof:** To verify this formula, let us assume we already know the controllable companion form, $(\mathbf{A}_c, \mathbf{B}_c)$ of the given controllable pair $(\mathbf{A}, \mathbf{B})$. We've already shown that

$$\mathbf{F}_c = \begin{bmatrix} a_0 + d_0 & \cdots & a_{n-1} + d_{n-1} \end{bmatrix}$$

where the characteristic polynomial of $\mathbf{A}$ is

$$p_A(s) = s^n + a_{n-1}s^{n-1} + \cdots + a_1 s + a_0$$

The matrix $\mathbf{F}$ is related to $\mathbf{F}_c$ through the similarity transformation $\mathbf{F} = \mathbf{F}_c\mathbf{P}$ where $\mathbf{P} = \mathcal{C}_c\mathcal{C}^{-1}$ and

$$\mathcal{C}_c = \text{controllability matrix if } (\mathbf{A}_c, \mathbf{B}_c)$$

We will now show that

$$\mathbf{F}_c = -e_n^T \mathcal{C}_c^T \alpha_d(\mathbf{A}_c)$$

arbitarily assigns the $n$ eigevalues of $\mathbf{A}_c$ to the roots of the desired plynomial $\alpha_d(s)$. Note that

$$\alpha_d(\mathbf{A}_c) = \mathbf{A}_c^n + d_{n-1}\mathbf{A}_c^{n-1} + \cdots + d_1\mathbf{A}_c + d_0\mathbf{I}$$

through the Cayley- Hamilton theorem we know that

$$p_A(\mathbf{A}_c) = \mathbf{A}_c^n + a_{n-1}\mathbf{A}_c^{n-1} + \cdots + a_1\mathbf{A}_c + a_0\mathbf{I} = 0$$

which means that

$$\alpha_d(\mathbf{A}_c) = \sum_{i=0}^{n-1}(a_i + d_i)\mathbf{A}_c^i$$

Also note that

$$\begin{aligned} e_1^T\mathcal{C}_c &= \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & \cdots & 1 \\ \vdots & \vdots & & \vdots \\ 0 & 1 & \cdots & x \\ 1 & x & \cdots & x \end{bmatrix} \\ &= e_n^T \end{aligned}$$

or rather that $e_n^T\mathcal{C}_c^{-1} = e_1^T$. Pre-multiply $\alpha_d(\mathbf{A}_c)$ by $-e_n^T\mathcal{C}_c^{-1}$ to get

$$\begin{aligned} -e_n^T\mathcal{C}_c^{-1}\alpha_d(\mathbf{A}_c) &= -e_1^T\left[(d_0 + a_0)\mathbf{I} + \cdots + (d_{n-1} + a_{n-1})\mathbf{A}_c^{n-1}\right] \\ &= \begin{bmatrix} a_0 + d_0 & a_1 + d_1 & \cdots & a_{n-1} + d_{n-1} \end{bmatrix} \\ &= \mathbf{F}_c \end{aligned}$$

which verifies Ackerman's formula when the system is in its controllable companion form.

To complete the proof, simply transform $\mathbf{F}_c$ back to the original realization using the similarity transformation $\mathbf{P} = \mathcal{C}_c\mathcal{C}^{-1}$. This gives

$$\begin{aligned}
\mathbf{F} &= \mathbf{F}_c\mathbf{P} = -e_n^T\mathcal{C}_c^{-1}\alpha_d(\mathbf{A}_c)\mathbf{P} \\
&= -e_n^T\mathcal{C}_c^{-1}\alpha_d(\mathbf{P}\mathbf{A}\mathbf{P}^{-1})\mathbf{P} = -e_n^T\mathcal{C}_c^{-1}\mathbf{P}\alpha_d(\mathbf{A}) \\
&= -e_n^T\mathcal{C}_c^{-1}(\mathcal{C}_c\mathcal{C}^{-1})\alpha_d(\mathbf{A}) \\
&= -e_n^T\mathcal{C}^{-1}\alpha_d(\mathbf{A})
\end{aligned}$$

which completes the proof. $\diamondsuit$

To illustrate the use of this method, let us return the earlier example which had the state space realization,

$$\mathbf{G} \stackrel{s}{=} \left[\begin{array}{ccc|c} \frac{1}{3} & \frac{1}{3} & -\frac{2}{3} & \frac{1}{3} \\ \frac{1}{3} & -\frac{2}{3} & \frac{1}{3} & \frac{1}{3} \\ -\frac{2}{3} & -\frac{5}{3} & -\frac{2}{3} & \frac{1}{3} \\ \hline 1 & 1 & -2 & 0 \end{array}\right]$$

The desired polynomial was

$$\alpha_d(s) = (s + 1/2)(s^2 + 3s + 2) = s^3 + 3.5s^2 + 3.5s + 1$$

and so the following script will compute the desired gain $\mathbf{F}$ directly without having to first convert the realization to its controller companion form.

```
en = zeros(3,1); en(3)=1;
adA = A^3+3.5*A^2+3.5*A+eye(3,3);
F     = -en'*inv(ctrb(A,B))*adA
```

which gives the output

```
F =

 -4.5000   -4.0000    1.0000
```

which is identical to what we computed before.

**Remark:** We used the MATLAB command `ctrb` to compute the controllability matrix. The computations in Ackerman's formula are easily encapsulated into a MATLAB function. In particular, MATLAB has already done this using the command `acker` which is in MATLAB's control system toolbox.

## 2. Luenberger Observer

In many applications, one may not have direct access to the full state. In this case, one would like to find a way to *estimate* the full state, $x$, from the available observed outputs, $y$. Such a system is called an *observer*. This section discusses methods used for building such observers. The tools we developed above to identify stabilizing state feedback laws will be used to determine such observers.

First let us consider an "open-loop" way of constructing a state observer for a plant, $\mathbf{G}$, with state space realization

$$\begin{aligned} \dot{x}(t) &= \mathbf{A}x(t) + \mathbf{B}_1 w(t) + \mathbf{B}_2 u(t) \\ y(t) &= \mathbf{C}x(t) + \mathbf{D}u(t) \end{aligned} \tag{43}$$

where $x : \mathbb{R} \to \mathbb{R}^n$ is the system state. We supply the other signals with the following "physical" interpretations

$$\begin{aligned} u : \mathbb{R} \to \mathbb{R} \text{ is a "known" control input} \\ y : \mathbb{R} \to \mathbb{R} \text{ is a "observed" system output} \\ w : \mathbb{R} \to \mathbb{R} \text{ is an "unknown" external disturbance} \end{aligned}$$

We assume that the state space realization for $\mathbf{G} \stackrel{s}{=} \left[\begin{array}{c|cc} \mathbf{A} & \mathbf{B}_1 & \mathbf{B}_2 \\ \hline \mathbf{C} & \mathbf{0} & \mathbf{D} \end{array}\right]$ is known. Note that the "$\mathbf{B}$" and "$\mathbf{D}$" matrices have a block form determined by the two types of inputs, $u$ and $w$, that are driving the system.

Since we know the system matrices, we can try to build a state estimator that simply "mimics" the dynamics of the plant. If we let $\widehat{x} : \mathbb{R} \to \mathbb{R}^n$ denote the estimated state, then this would mean we might try an estimator of the form

$$\begin{aligned} \dot{\widehat{x}}(t) &= \mathbf{A}\widehat{x}(t) + \mathbf{B}_2 u(t) \\ \widehat{y}(t) &= \mathbf{C}\widehat{x}(t) + \mathbf{D}u(t) \end{aligned} \tag{44}$$

Note that we have not included the terms determined by the external disturbance $w$ since this signal is not known. So the estimator is an LTI system with input $u$ (which is known) and an output $\widehat{y}$ which is the "estimated" output based on the information in $u$ and $\widehat{x}$.

We want to study how the state estimation error,

$$\widetilde{x}(t) \stackrel{\text{def}}{=} x(t) - \widehat{x}(t)$$

behaves. In particular, $\widetilde{x}$, should satisfy the following differential equation

$$\begin{aligned} \dot{\widetilde{x}}(t) &= \dot{x}(t) - \dot{\widehat{x}}(t) \\ &= \mathbf{A}x(t) + \mathbf{B}_1 w(t) + \mathbf{B}_2 u(t) - \mathbf{A}\widehat{x}(t) - \mathbf{B}_2 u(t) \\ &= \mathbf{A}\widetilde{x}(t) + \mathbf{B}_1 w(t) \end{aligned}$$

Note that if $\mathbf{A}$'s eigenvalues all have negative real parts, then the origin of the unforced system when $w = 0$ is asymptotically stable and so the estimation error would go to zero asymptotically. Let $\mathbf{H}(s)$ denote the transfer function from $w$ to $\widetilde{x}$. One can readily see that

$$\mathbf{H}(s) = (s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}_1$$

If $w$ is an $\mathcal{L}_2$ signal, then we know the $\mathcal{L}_2$ norm of the state estimation error is bounded by

$$\|\widetilde{x}\|_{\mathcal{L}_2} \leq \|\mathbf{H}\|_{\mathcal{H}_\infty} \|w\|_{\mathcal{L}_2}$$

which allows us to show that the norm of the estimation error is bounded above. Note that this bound again only exists if the system is finite gain $\mathcal{L}_2$ stable, which will always be the case if $\mathbf{A}$ is Hurwitz.

What the preceding discussion showed is that we get useful estimates of the state using the "open-loop" observer in equation (44) whenever $\mathbf{A}$ is Hurwitz, i.e. the origin of the open-loop system (43) is asymptotically stable. However, we would also like to build observers for *unstable* plants since we will later try to use those estimated states in a state-feedback law to stabilize the system. If we try our open-loop strategy on an unstable plant, then the state estimates will become unbounded thereby leaving us with a useless estimate of the state. To address this issue, we modify the observer equation to use the *observation error*

$$\widetilde{y}(t) \stackrel{\text{def}}{=} y(t) - \widehat{y}(t)$$

as a state feedback term, just as it was done above with the state feedback control law. The estimated output is taken to be

$$\widehat{y}(t) = \mathbf{C}\widehat{x}(t) + \mathbf{D}u(t)$$

The resulting *closed-loop* estimator is now called an *observer* and is defined by the following state equations

$$\begin{aligned} \dot{\widehat{x}}(t) &= \mathbf{A}\widehat{x}(t) + \mathbf{B}_2 u(t) + \mathbf{L}(y(t) - \widehat{y}(t)) \\ \widehat{y}(t) &= \mathbf{C}\widehat{x}(t) + \mathbf{D}u(t) \end{aligned} \tag{45}$$

where $\mathbf{L} \in \mathbb{R}^n$ is a matrix of *observer gains*. These observer gains must be chosen by the designer to ensure that the estimation error, $\widetilde{x}(t)$, asymptotically goes to zero in the absence of any external disturbance, $w$ and remains sufficiently bounded when $w$ is a bounded disturbance.

To determine the conditions needed for a "stable" observer, we write out the state equations for the state estimation error, $\widetilde{x}$,

$$\begin{aligned} \dot{\widetilde{x}}(t) &= \dot{x}(t) - \dot{\widehat{x}}(t) \\ &= \mathbf{A}x(t) + \mathbf{B}_1 w(t) + \mathbf{B}_2 u(t) \\ &\quad -\mathbf{A}\widehat{x}(t) - \mathbf{B}_2 u(t) - \mathbf{L}(\mathbf{C}x(t) + \mathbf{D}u(t) - \mathbf{C}\widehat{x}(t) - \mathbf{D}u(t)) \\ &= \mathbf{A}\widetilde{x}(t) - \mathbf{L}\mathbf{C}\widetilde{x}(t) + \mathbf{B}_1 w(t) \\ &= (\mathbf{A} - \mathbf{L}\mathbf{C})\widetilde{x}(t) + \mathbf{B}_1 w(t) \end{aligned} \tag{46}$$

We obviously want to choose $\mathbf{L}$ so that all of the eigenvalues of $\mathbf{A} - \mathbf{LC}$ have negative real parts. If this is done then the estimation error goes to zero asymptotically when $w = 0$ and remains bounded in $\mathcal{L}_2$ if $w$ is also $\mathcal{L}_2$.

We call this observer in equation (45) a *Luenberger observer*. Note that choosing $\mathbf{L}$ is equivalent to a pole placement problem. In particular, we can see that $\mathbf{A} - \mathbf{LC}$ and $(\mathbf{A} - \mathbf{LC})^T$ have the same eigenvalues. Note that

$$(\mathbf{A} - \mathbf{LC})^T = \mathbf{A}^T - \mathbf{C}^T\mathbf{L}^T$$

So our ability to freely assign the eigenvalues of $\mathbf{A}-\mathbf{LC}$ is equivalent to the pair $(\mathbf{A}^T, \mathbf{C}^T)$ being arbitrarily assignable. We know this is the case if and only if $(\mathbf{A}^T, \mathbf{C}^T)$ is controllable which means that $(\mathbf{A} - \mathbf{LC})$'s eigenvalues can be freely assigned if and only if $(\mathbf{A}, \mathbf{C})$ is observable. We can therefore summarize our preceding discussion in the following theorem.

THEOREM 39. *The eigenvalues of the Luenberger observer in equation (45) can be freely assigned if and only if* $(\mathbf{A}, \mathbf{C})$ *is observable.*

Clearly, we want to choose $\mathbf{L}$ so that $\mathbf{A} - \mathbf{LC}$ is Hurwitz. When this is the case then we say the pair $(\mathbf{A}, \mathbf{C})$ be *detectable*. A necessary and sufficient condition for detectability is that all unobservable eigenvalues of $(\mathbf{A}, \mathbf{C})$ have negative real parts.

Recall that asymptotic stability of the origin can only be assured if the external disturbance, $w(t) = 0$. When $w$ is not zero, then we require the input/output system from $w$ to the state estimation error $\widetilde{x}$ be finite gain $\mathcal{L}_2$ stable. Let $\mathbf{E}$ denote the error system in equation (46) whose state space realization is

$$\mathbf{E} \stackrel{s}{=} \left[\begin{array}{c|c} \mathbf{A} - \mathbf{LC} & \mathbf{B}_1 \\ \hline \mathbf{I} & \mathbf{0} \end{array}\right]$$

We are going to choose $\mathbf{L}$ so the origin of the unforced error system is asymptotically stable. We know that this will also imply the input/output system, $\mathbf{E}$, is $\mathcal{L}_2$ stable. In this case, we will use the $\mathcal{L}_2$-induced gain of $\mathbf{E}$ to characterize the "performance" level achieved by the estimator; with

smaller gains being associated with better performance. In our earlier work we showed that the $\mathcal{L}_2$ induced gain is given by the $\mathcal{H}_\infty$ norm of the transfer function matrix', $\mathbf{E}(s)$, maximum gain magnitude.

$$\|\mathbf{E}\|_{\mathcal{H}_\infty} = \sup_\omega |\mathbf{E}(j\omega)|$$

which we can evaluate by an examination of the error system's gain magnitude plot, or through the use of the bounded real lemma in a bisection search. There are certain applications where this approach is used to select gains, $\mathbf{L}$, that *minimize* the $\mathcal{H}_\infty$ gain of $\mathbf{E}$, thereby maximizing the estimator's performance. The resulting "optimal" $\mathcal{H}_\infty$ filters are often used when we want the estimator's performance to be *robust* to model uncertainty.

There are a large number of applications where the input $w$ is a unit variance white noise process. In this case the state estimate $\{\widetilde{x}(t)\}$ is a stochastic process that is normally distributed so we only need to determine its mean and covariance matrix. In particular, let $\mathbb{E}\{\widetilde{x}(t)\} = \mu(t)$, then because of the linearity of the expectation operator and because $w$ is zero mean, we can see

$$\frac{d}{dt}\mathbb{E}\{\widetilde{x}(t)\} = \dot{\mu}(t) = (\mathbf{A} - \mathbf{LC})\mathbb{E}\{\widetilde{x}(t)\} + \mathbf{B}_1\mathbb{E}\{w(t)\} = (\mathbf{A} - \mathbf{LC})\,\mu(t)$$

If $\mathbf{L}$ is chosen so $(\mathbf{A} - \mathbf{LC})$ is Hurwitz then we can see that $\mu(t) \to 0$ as $t \to \infty$ for any particular input $w$ that we have.

A full characterization of the state estimation process, however, also requires us to determine its covariance matrix $\mathbf{E}\{\widetilde{x}(t)\widetilde{x}^T(t)\} = \mathbf{P}(t)$. Determination of the differential equation satisfied by the covariance matrix requires that we use methods from the Ito stochastic calculus [Fleming and Rishel (1972),Karatzas and Shreve (1998)]. Since these methods are beyond the scope of this course, we will simply summarize the results. In particular, one can use the Ito calculus to show that $\mathbf{P}(t)$ satisfies the following ordinary matrix differential equation

$$\dot{\mathbf{P}}(t) = (\mathbf{A} - \mathbf{LC})\mathbf{P}(t) + \mathbf{P}(t)(\mathbf{A} - \mathbf{LC})^T + \mathbf{B}_1\mathbf{B}_1^T$$

When $\mathbf{A}-\mathbf{LC}$ is Hurwitz then as $t \to \infty$ one can show that $\mathbf{P}(t)$ converges asymptotically to a constant matrix, $\mathbf{P}$

$$\lim_{t\to\infty} \mathbb{E}\left\{\widetilde{x}(t)\widetilde{x}^T(t)\right\} = \mathbf{P}$$

that satisfies the following Lyapunov equation

$$0 = (\mathbf{A}-\mathbf{LC})\mathbf{P} + \mathbf{P}(\mathbf{A}-\mathbf{LC})^T + \mathbf{B}_1\mathbf{B}_1^T \tag{47}$$

So in terms of characterizing the steady-state performance of the estimator, we would simply need to solve equation (47) for $\mathbf{P}$ whose trace could be used as a single number measuring the "performance" of the estimator. Finding an $\mathbf{L}$ that minimizes the $\mathrm{trace}(\mathbf{P})$ yields an "optimal" observer that minimizes the steady-state mean squared estimation estimation. The resulting observer is more commonly known as the steady state *Kalman filter* and the equations characterizing this optimal $\mathbf{L}$ will be discussed below.

## 3. Linear Quadratic Regulator

The preceding section showed how state feedback can be used to freely assign the eigenvalues of an LTI system. The next question, of course, is why one would want to do this and where are the "best" locations for these eigenvalues. Rather than answering this directly, we will pose the question as an *optimization problem* that seeks a state feedback gain matrix, $\mathbf{F}$, that optimizes some useful measure of how we believe the system should perform. Consider the state space equation

$$\begin{aligned} \dot{x} &= \mathbf{A}x + \mathbf{B}_1 w + \mathbf{B}_2 u \\ y &= x \\ z &= \begin{bmatrix} \mathbf{C}x \\ u \end{bmatrix} \end{aligned}$$

with initial condition $x(0) = x_0$. The signal, $y$, is our usual output signal which, in this case, provides full access to the system state. We refer to this as the *Full Information* or FI controller. The additional output signal, $z$, is a signal used to characterize the performance of the system, with smaller $z$

meaning better performing systems. There are two inputs; $w$ and $u$. The input $w$ is a disturbance signal that we take to be a unit variance white noise process. The other input $u$ is the control input to be supplied by a controller we need to design. Since we have full state access, we know our control signal will be $u = \mathbf{F}x$. The problem is how do we go about selecting $\mathbf{F}$ in an "optimal" manner? Optimality depends on how we wish to characterize performance, but if we take $z$ as a "virtual" signal used to help use measure system performance, then we will want to select $\mathbf{F}$ to minimize

$$\mathbb{E}\left[\|z\|^2_{\mathcal{L}_2}\right]$$

Namely we select a gain that minimized the mean squared energy in the objective signal.

We first consider a *finite horizon* version of this optimization problem that seeks to minimize the following cost functional,

$$\begin{aligned} J[u;T,\mathbf{M}] &\equiv \|z\|^2_{\mathcal{L}_2[0,T]} + x^T(T)\mathbf{M}x(T) \\ &= \int_0^T z^T(\tau)z(\tau)d\tau + x^T(T)\mathbf{M}x(T) \end{aligned}$$

where $\mathbf{M}$ is a symmetric positive definite matrix. The first term on the right-hand side represents a path (running) cost and the final term is a terminal cost associated with not zeroing the output by time $T$. This section finds a $u$ that minimizes this finite horizon cost and then examines how it behaves as $T$ goes to infinity.

We start by assuming a solution exists and then identify necessary conditions that the solution must satisfy. These necessary conditions then provide the means for determining the optimal control gain $\mathbf{F}$. Let us first introduce a matrix-valued function, $\mathbf{X} : \mathbb{R} \to \mathbb{R}^{n\times n}$ where $\mathbf{X}(t) \geq 0$ for all $t \in [0,T]$ with $\mathbf{X}(T) = \mathbf{M}$. We may then rewrite the cost functional, $J$, as

$$\begin{aligned} J[u\,;\,T,\mathbf{M}] &= \int_0^T z^T(\tau)z(\tau)d\tau + x^T(T)\mathbf{M}x(T) \\ &= \int_0^T \left(z^T(\tau)z(\tau) + \frac{d}{d\tau}x^T(\tau)\mathbf{X}(\tau)x(\tau)\right)d\tau + x^T(0)\mathbf{X}(0)x(0) \end{aligned}$$

Note that

$$\frac{d}{dt}(x^T\mathbf{X}x) = \dot{x}^T\mathbf{X}x + x^T\mathbf{X}\dot{x} + x^T\dot{\mathbf{X}}x$$

Inserting this into the preceding expression for $J[u\,;\,T,\mathbf{M}]$ yields,

$$\begin{aligned} J[u\,;\,T,\mathbf{M}] &= \int_0^T \left(x^T(\tau)\mathbf{C}^T\mathbf{C}x(\tau) + u^T(\tau)u(\tau)\right) d\tau \\ &\quad + \int_0^T \left[2(x^T\mathbf{A}^T + w^T\mathbf{B}_1^T + u^T\mathbf{B}_2^T)\mathbf{X}x + x^T\dot{\mathbf{X}}x\right] d\tau \\ &\quad + x^T(0)\mathbf{X}(0)x(0) \end{aligned}$$

Collecting the quadratic terms in $x$ yields

$$\begin{aligned} J[u\,;\,T,\mathbf{M}] &= \int_0^T x^T(\mathbf{C}^T\mathbf{C} + \mathbf{A}^T\mathbf{X} + \mathbf{X}\mathbf{A} + \dot{\mathbf{X}})x d\tau \\ &\quad + \int_0^T \left(u^T u + 2u^T\mathbf{B}_2^T\mathbf{X}x + 2w^T\mathbf{B}_1^T\mathbf{X}x\right) d\tau \\ &\quad + x^T(0)\mathbf{X}(0)x(0) \end{aligned}$$

We complete the square of the first two terms on the second line by adding and subtracting the term $x^T\mathbf{X}\mathbf{B}_2^T\mathbf{B}_2\mathbf{X}x$. This lets us rewrite $J$ as

$$\begin{aligned} J[u\,;\,T,\mathbf{M}] &= \int_0^T x^T\left(\mathbf{C}^T\mathbf{C} + \mathbf{A}^T\mathbf{X} + \mathbf{X}\mathbf{A} + \dot{\mathbf{X}} - \mathbf{X}\mathbf{B}_2\mathbf{B}_2^T\mathbf{X}\right) x d\tau \\ &\quad + \int_0^T \left(\left|u + \mathbf{B}_2^T\mathbf{X}x\right|^2 + 2w^T\mathbf{B}_1^T\mathbf{X}x\right) d\tau + x^T(0)\mathbf{X}(0)x(0) \end{aligned}$$

If we select $u^* = -\mathbf{B}_2^T\mathbf{X}x$ and we let $\mathbf{X}$ satisfy the following *matrix differential Riccati equation*

$$\begin{aligned} -\dot{\mathbf{X}} &= \mathbf{A}^T\mathbf{X} + \mathbf{X}\mathbf{A} - \mathbf{X}\mathbf{B}_2\mathbf{B}_2^T\mathbf{X} + \mathbf{C}^T\mathbf{C} \\ \mathbf{X}(T) &= \mathbf{M} \end{aligned}$$

then the first two terms vanish and we have

$$J[u^*\,;\,T,\mathbf{M}] = \int_0^T 2w^T\mathbf{B}_1^T\mathbf{X}x d\tau + x^T(0)\mathbf{X}(0)x(0)$$

Since $w$ is a zero mean white noise process that is statistically independent from $x$, then if we take the expectation of $J$ we get

$$\begin{aligned}\mathbb{E}\left[J[u^*\,;\,T,\mathbf{M}]\right] &= \mathbf{E}\left[\int_0^T 2w^T\mathbf{B}_1^T\mathbf{X}x d\tau\right] + x^T(0)\mathbf{X}(0)x(0)\\ &= x^T(0)\mathbf{X}(0)x(0)\end{aligned}$$

The first term vanishes because $x$ and $w$ are statistically independent. So the optimal cost is simply $x^T(0)\mathbf{X}(0)x(0)$ where $x(0)$ is the initial state and $\mathbf{X}(0)$ is obtained by solving the Riccati differential matrix equation backward in time from $T$.

To summarize the finite horizon problem is solved by

$$u^* = -\mathbf{B}_2^T\mathbf{X}(t)x(t)$$

where

$$\begin{aligned}-\dot{\mathbf{X}} &= \mathbf{A}^T\mathbf{X} + \mathbf{X}\mathbf{A} - \mathbf{X}\mathbf{B}_2\mathbf{B}_2^T\mathbf{X} + \mathbf{C}^T\mathbf{C}\\ \mathbf{X}(T) &= \mathbf{M}\end{aligned}$$

This has a solution provided $\mathbf{C}^T\mathbf{C} > 0$.

If we let $T$ go to infinity it can be shown [Green and Limebeer (2012)] that if we choose $\mathbf{M} \geq 0$ such that

$$\mathbf{M}\mathbf{A} + \mathbf{A}^T\mathbf{M} - \mathbf{M}\mathbf{B}_2\mathbf{B}_2^T\mathbf{M} + \mathbf{C}^T\mathbf{C} \leq 0$$

then

$$\lim_{T\to\infty} \mathbf{X}(t;T,\mathbf{M}) = \mathbf{\Pi} = \text{constant matrix}$$

This constant matrix must satisfy the algebraic Riccati equation

$$0 = \mathbf{A}^T\mathbf{\Pi} + \mathbf{\Pi}\mathbf{A} - \mathbf{\Pi}\mathbf{B}_2\mathbf{B}_2^T\mathbf{\Pi} + \mathbf{C}^T\mathbf{C}$$

which is the standard LQR Riccati equation. The control is "optimal" in the sense of minimizing the expected value of $J[u]$ and this control is given by

$$u^* = -\mathbf{B}_2^T\mathbf{\Pi}x$$

The matrix $-\mathbf{B}_2^T\mathbf{\Pi}$ corresponds to the gain matrix, $\mathbf{F}$, we discussed earlier when studying pole placement methods. It can be further shown [Green and Limebeer (2012)] that the origin of this controlled system is asymptotically stable assuming that $\left(\mathbf{A}, \begin{bmatrix} \mathbf{C} \\ \mathbf{M} \end{bmatrix}\right)$ is detectable.

**Example:** Let us now give an example to illustrate the use of the LQR regulator. In this case we consider the servo positioning system shown in Fig. 1. The motor is used to move a load in a translational manner. The angular position, $\theta$, of the motor determines the position of the load. If we ignore the electrical part of the motor and only focus on the mechanical states, then the servo's dynamics may be modeled as a double integrator that is driven by the applied torque $u$ generated by the motor, a disturbance torque, $w$, that comes from the load, and a damping torque, $\alpha\dot{\theta}$. The servo position therefore satisfies the following ODE,

$$\ddot{\theta} = u + w + \alpha\dot{\theta}$$

where $\alpha$ is a known positive constant.

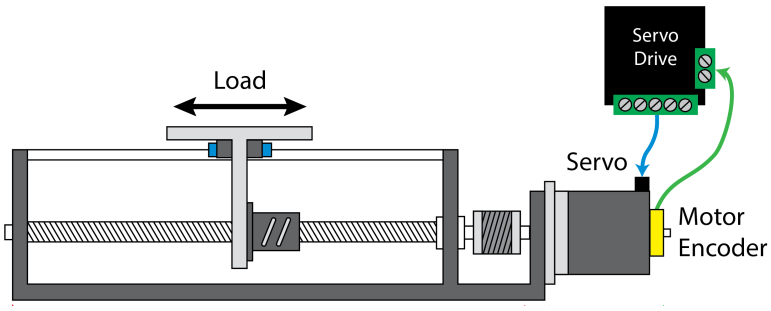



FIGURE 1. Servo Positioning System

To find the LQR controller, we first need to rewrite the preceding second order ODE as a pair of first order differential equations. We first introduce

the following state variable

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \theta - \theta_c \\ \dot{\theta} \end{bmatrix}$$

The state equations governing $x$ are

$$\begin{aligned} \dot{x} &= \begin{bmatrix} \dot{\theta} - \dot{\theta}_c \\ \ddot{\theta} \end{bmatrix} \\ &= \begin{bmatrix} x_2 \\ -\alpha x_2 + u(t) + w(t) \end{bmatrix} \\ &= \begin{bmatrix} 0 & 1 \\ 0 & -\alpha \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} (u(t) + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \end{aligned}$$

We can now see that in the LQR formalism our system matrices are

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ 0 & -\alpha \end{bmatrix}, \quad \mathbf{B}_1 = \mathbf{B}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

and our LQR cost functional takes the form

$$J[u] = \int_0^\infty (qx_1^2 + u^2)dt = \int_0^\infty (x^T \mathbf{Q} x + u^2)$$

where $\mathbf{Q} = \begin{bmatrix} q & 0 \\ 0 & 0 \end{bmatrix}$ and $\mathbf{R} = 1$.

A MATLAB script was written to compute the LQR gains and simulate the resulting system. The results from the simulation are shown in Fig. 2. In this simulation, $\theta_c = -1$ and the initial servo angle, $\theta(0) = 0$, and an initial servo rate $\dot{\theta}(0) = 1$. The damping ratio $\alpha$ was set to $0.1$. Figure 2 shows the servo angle, $\theta$, (top plot) and the servo rate, $\dot{\theta}$, and control $u$ (bottom plot) for $q = .1$, $1$, and $10$. A small amount of process noise $w$ was added into the simulation results also. As the weighting parameter $q$ increases, we expect a higher penalty on large deviations away from zero. This means that for larger $q$, we expect $\theta(t)$ to converge more quickly to $\theta_c$ with a smaller variation. This is exactly what is seen in the top plots. For larger $q$, however, we also expect a smaller penalty to be paid for larger

control effort and so for larger $q$ we expect to see larger control torques $u$. This is also seen in the bottom plots of Fig. 2

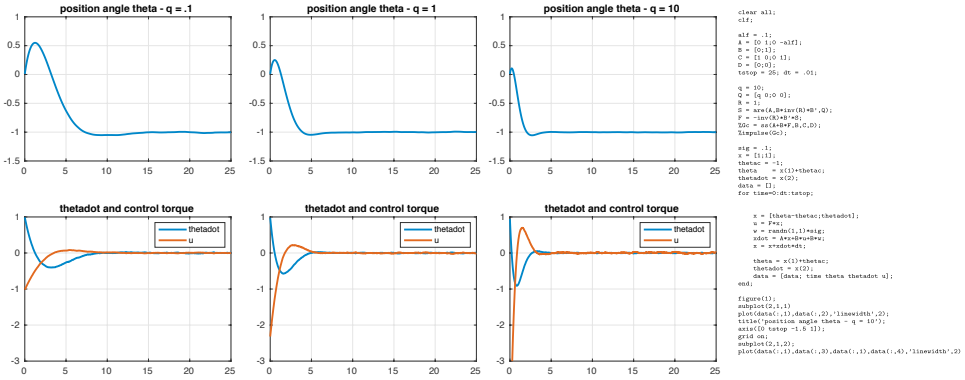



FIGURE 2. Simulations of LQR controlled servo

One important thing to notice here is that the LQR control law takes the form

$$u = f_1(\theta - \theta_c) + f_2(\dot{\theta})$$

From Fig. 1, however, we see that the only sensor on the servo is an encoder measuring the shaft position. In other words, we don't have the angular rate of the servo available for the control. This means, of course, that this particular control system is not "implementable". Finding a way to physically implement this control will be the subject of the next sections.

## 4. Steady State Kalman Filter

By duality [Green and Limebeer (2012)], one can develop optimal full state observers. We will not do that here and will simply state the main result. The optimal state observer is called a steady-state *Kalman filter* and seeks to minimize the MSE $\lim_{t\to\infty} \mathbb{E}\{z^T(t)z(t)\}$.

We take the open loop plant to have the form,

$$\begin{aligned} \dot{x} &= \mathbf{A}x + \mathbf{B}_1 w + \mathbf{B}_2 u \\ y &= \mathbf{C}x + v \end{aligned}$$

where $w$ and $v$ are uncorrelated white noise processes. The other signal $u$ is a control signal which we are assumed to know. The state equation for both filters are

$$\dot{\hat{x}} = \left(\mathbf{A} - \mathbf{Q}\mathbf{C}^T\mathbf{C}\right)\hat{x} + \mathbf{B}_2 u + \mathbf{Q}\mathbf{C}^T y$$

where $\mathbf{Q} = \mathbf{Q}^T > 0$ satisfies an algebraic Riccati equation. For the $\mathcal{H}_2$ (i.e. steady state Kalman filter) this ARE is

$$0 \quad = \quad \mathbf{Q}\mathbf{A}^T + \mathbf{A}\mathbf{Q} - \mathbf{Q}\mathbf{C}^T\mathbf{C}\mathbf{Q} + \mathbf{B}_1\mathbf{B}_1^T$$

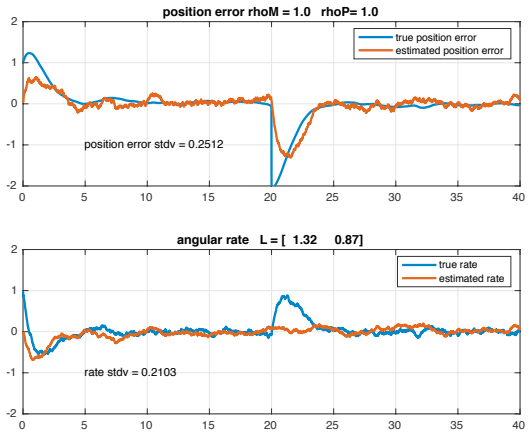
Note that the above equation for the observer can be rewritten to look like a Luenberger observer; thereby emphasizing the feedback aspect of the observer. In particular, since $\hat{y} = \mathbf{C}\hat{x}$, it should be apparent that

$$\begin{aligned} \dot{\hat{x}} &= \mathbf{A}\hat{x} + \mathbf{B}_2 u + \mathbf{Q}\mathbf{C}^T(y - \hat{y}) \\ \hat{y} &= \mathbf{C}\hat{x} \end{aligned}$$

In this case we see that the optimal observer gains are

$$\mathbf{L}_{\text{optimal}} = \mathbf{Q}\mathbf{C}^T$$

Let us see how well this filter works on the earlier LQR controlled servo system. In this case, we use the LQR gains computed when $q = 1$ and assume that the process noise and measurement noise both have unit covariance. The simulation starts with $\theta_c = -1$ at $t = 0$ and then switches to $\theta_c = 1$ at $t = 20$. Fig. 3 plots the true state and estimated states as a function of time. The top plot is for the position error $\theta - \theta_c$ and the bottom plot is for the angular rate $\dot{\theta}_c$. Both plots show that the filter estimates track the actual states. We see that the position error estimate responds more quickly to a step change than the angular rate estimate, which is to be expected since we have a direct measurement of the position error. Our estimate for the angular rate must rely on the past rate estimates, since we don't directly observe the motor's rate.




```
clear all;
clf;

rhoP=1;
rhoM=10;
alf = .1;
A = [0 1;0 -alf];
B = [0;1];
C = [1 0];
D = [0];
tstop = 40; dt = .01;

q = 1;Q = q*eye(2,2);R = 1;
S = are(A,B*inv(R)*B',Q);
F = -inv(R)*B'*S;

W = eye(1,1)+rhoP;V = eye(1,1)+rhoM;
P = are(A',C'*inv(V)*C,B*W*B');
L = P*C'*inv(V);

sig = 1;x = [1;1];xhat = [0;0];
thetac = -1;theta = x(1)+thetac;
thetadot = x(2);
data = [];Psamp=0;
for time=0:dt:tstop;
    if time >=20;
        thetac = 1;
    end;
    w = randn(1,1)*sig;
    v = randn(1,1)*sig;
    x = [theta-thetac;thetadot];
    y = C*x+v;
    u = -(theta-thetac);
    u = F*x;
    xdot = A*x+B*u+B*w;
    xhatdot = A*xhat+B*u+L*(y-C*xhat);
    x = x+xdot*dt;
    xhat = xhat+xhatdot*dt;
    xtilde = x-xhat;
    theta = x(1)+thetac;
    thetadot = x(2);
    data = [data; time x' xhat'];
    Psamp = Psamp + (xtilde*xtilde')/4000;
end;
```

FIGURE 3. Kalman Filter simulation - Model matches noise covariances

As mentioned above, the Kalman filter minimizes the mean square estimation error. But that guarantee of optimality requires that we were truthful in telling the Riccati equation about the system matrices and the noise covariances. Of particular concern to us are the noise covariances, for it may be very difficult in practice to know exactly what these covariances might be ahead of time. In practice this means that one may often design a filter using only a best "guess" about the noise covariances. Such guesses can greatly impact the filter's performance.

To explore this sensitivity, let us assume the actual measurement and process noise in the system have unit variance, but that we designed the Kalman filter assuming $\mathbf{V} = 1$ (accurate) and $\mathbf{W}$ being $0.1$ or $10$ (too small or too big by an order of magnitude). If the filter were designed with $\mathbf{W} = 0.1$, then we lied to the filter by saying there is less process noise than there really is. This causes the filter to trust its past estimates more than the new measurements and so the observer gains are smaller. We would also expect to see the position estimate to converge more slowly with larger fluctuations than the rate estimate. These predictions are borne out on the lefthand side of Fig. 4. On the other hand, if we had designed the filter with $\mathbf{W} = 10$, then we lied by saying there is more process noise than there really is. This

causes the filter to put more trust in new measurements than the past state estimates and so the observer gains are larger, the filter converges more quickly, and the rate estimates show greater fluctuations than the position estimates. These observations are also borne out by the plots on the right hand side of Fig. 4.

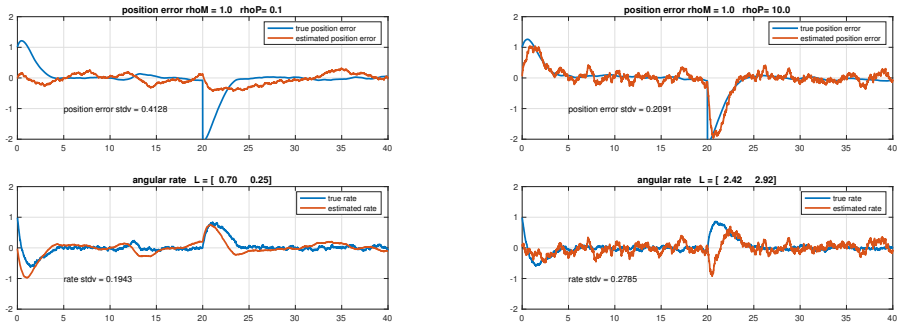



FIGURE 4. Kalman filter sensitivity to variations in Process Noise Covariance

We can repeat this experiment, but now look at how lying to the filter about the measurement noise covariance impacts the filter's performance. In this case we let $\mathbf{W} = 1$ and let $\mathbf{V} = 0.1$ or $10$. If we set $\mathbf{V} = 0.1$, then we are telling the filter there is less measurement noise than there really is and so the filter will trust its measurement more than the past state estimates. This means the observer gains will be larger, the position error will track the true error closely, and the rate estimate will have a larger error in it. This prediction is borne out on the left hand side of Fig. 5. On the other hand, if the filter was designed with $\mathbf{V} = 10$, then we are telling the filter there is more measurement noise than there really is and so the filter trusts its past state estimates more than the new measurements. This will be reflected in smaller observer gains and will suggest that the rate estimate will more closely track the position estimates. Again this prediction is borne out by the plots on the right hand side of Fig. 5.

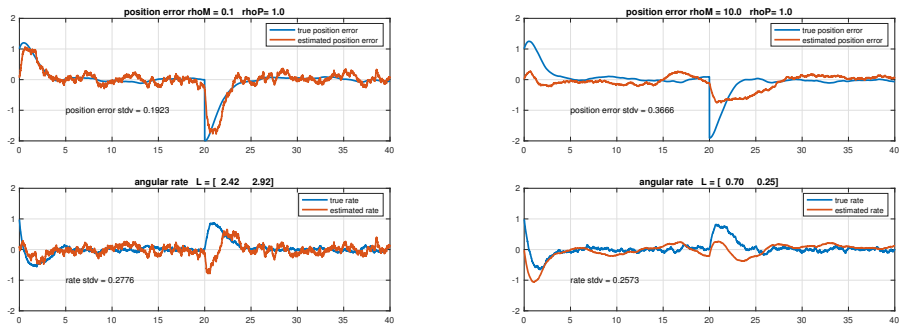


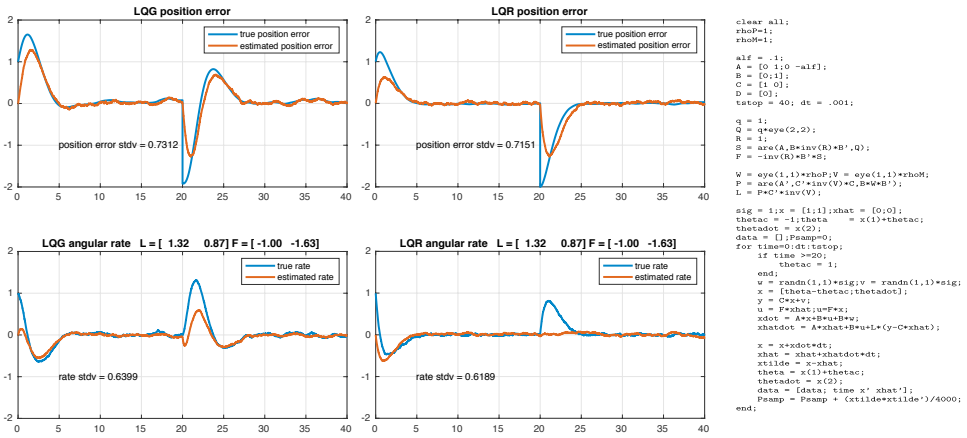

FIGURE 5. Kalman filter sensitivity to variations in Measurement Noise Covariance

The preceding empirical results show that the behavior of the Kalman filter is extremely sensitive to the prior information we had regarding the noise covariances. In general, these covariances are not exactly known. This is particular true of the process noise covariance, $w$. Process noise is really a term we use to approximate random inputs from the external environment as well as additive disturbances due to dynamics in the "physical" system that were neglected in our linear model of that system. The fact that we don't really know these covariances exactly means that designing Kalman filters is a recursive design procedure which often uses nonlinear simulations of the physical plant to empirically fine tune the Kalman filter.

## 5. Linear Quadratic Gaussian Controller

When the estimates generated by a Kalman filter are used in the LQR state feedback law, we obtain what is more commonly known as the *Linear Quadratic Gaussian* or LQG controller. One important feature of LQG controllers is that the performance achievable by the LQG controller is simply the sum of the performance of the LQR and Kalman filter. This fact is called the *separation principle* and it means that one can design the LQR and Kalman filter separately with an assurance that the composition of these two systems will still be optimal.

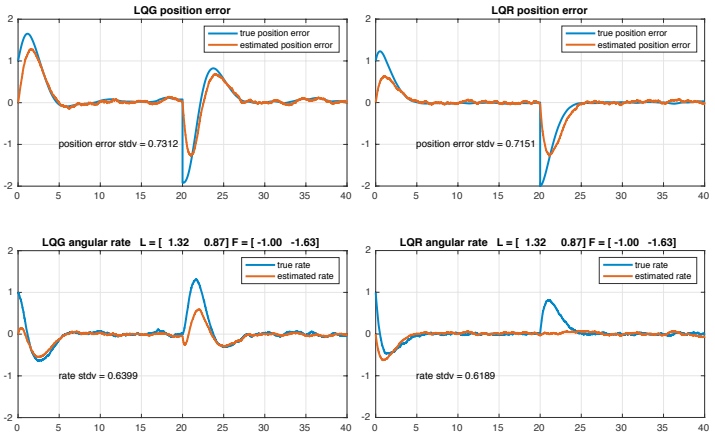
We now look at how well the servo system in Fig. 1 is regulated using the LQG controller. For this simulation we simply took the script used in Fig. 3 and replaced the line computing the control $u = Fx$ with $u = F\hat{x}$. The results for the LQG and associated LQR controller are shown side by side in Fig. 7. We see that both controllers are able to regulate the servo's position around the desired $\theta_c$. The difference lies in how well they do this. Since the LQG uses a noisy estimate of the state, rather than the true state, we see a larger regulation error, which is to be expected.




```
clear all;
rhoP=1;
rhoM=1;

alf = .1;
A = [0 1;0 -alf];
B = [0;1];
C = [1 0];
D = [0];
tstop = 40; dt = .001;

q = 1;
Q = q*eye(2,2);
R = 1;
S = are(A,B*inv(R)*B',Q);
F = -inv(R)*B'*S;

W = eye(1,1)*rhoP;V = eye(1,1)*rhoM;
P = are(A',C'*inv(V)*C,B*W*B');
L = P*C'*inv(V);

sig = 1;x = [1;1];xhat = [0;0];
thetac = -1;theta    = x(1)+thetac;
thetadot = x(2);
data = [];Psamp=0;
for time=0:dt:tstop;
    if time >=20;
       thetac = 1;
    end;
    v = randn(1,1)*sig;v = randn(1,1)*sig;
    x = [theta-thetac;thetadot];
    y = C*x+v;
    u = F*xhat;u=F*x;
    xdot = A*x+B*u+B*w;
    xhatdot = A*xhat+B*u+L*(y-C*xhat);

    x = x+xdot*dt;
    xhat = xhat+xhatdot*dt;
    xtilde = x-xhat;
    theta = x(1)+thetac;
    thetadot = x(2);
    data = [data; time x' xhat'];
    Psamp = Psamp + (xtilde*xtilde')/4000;
end;
```

FIGURE 6. (left) LQG controlled servo (right) LQR controlled servo

We now look at how well the servo system in Fig. 1 in chapter 1 is regulated using the LQG controller. For this simulation we simply took the script used in Fig. 3 and replaced the line computing the control $u = Fx$ with $u = F\widehat{x}$. The results for the LQG and associated LQR controller are shown side by side in Fig. 7. We see that both controllers are able to regulate the servo's position around the desired $\theta_c$. The difference lies in how well they do this. Since the LQG uses a noisy estimate of the state, rather than the true state, we see a larger regulation error, which is to be expected.




```
clear all;
rhoP=1;
rhoM=1;

alf = .1;
A = [0 1;0 -alf];
B = [0;1];
C = [1 0];
D = [0];
tstop = 40; dt = .001;

q = 1;
Q = q*eye(2,2);
R = 1;
S = are(A,B*inv(R)*B',Q);
F = -inv(R)*B'*S;

W = eye(1,1)*rhoP;V = eye(1,1)*rhoM;
P = are(A',C'*inv(V)*C,B*W*B');
L = P*C'*inv(V);

sig = 1;x = [1;1];xhat = [0;0];
thetac = -1;theta    = x(1)+thetac;
thetadot = x(2);
data = [];Psamp=0;
for time=0:dt:tstop;
    if time >=20;
        thetac = 1;
    end;
    w = randn(1,1)*sig;v = randn(1,1)*sig;
    z = [theta-thetac;thetadot];
    y = C*x+v;
    u = F*xhat;u=F*z;
    xdot = A*x+B*u+B*w;
    xhatdot = A*xhat+B*u+L*(y-C*xhat);

    x = x+xdot*dt;
    xhat = xhat+xhatdot*dt;
    xtilde = x-xhat;
    theta = x(1)+thetac;
    thetadot = x(2);
    data = [data; time x' xhat'];
    Psamp = Psamp + (xtilde*xtilde')/4000;
end;
```

FIGURE 7. (left) LQG controlled servo (right) LQR controlled servo

# Bibliography


Antsaklis, P. and Michel, A. N. (2006). *Linear systems*. Springer Science & Business Media.

Boyd, S., El Ghaoui, L., Feron, E., and Balakrishnan, V. (1994). *Linear matrix inequalities in system and control theory*. SIAM.

Fleming, W. H. and Rishel, R. W. (1972). *Deterministic and stochastic optimal control*. Springer-Verlag.

Green, M. and Limebeer, D. J. (2012). *Linear robust control*. Courier Corporation.

Hartman, P. (2002). *Ordinary differential equations*. Society for Industrial and Applied Mathematics, 2 edition.

Hespanha, J. P. (2018). *Linear systems theory*. Princeton university press.

Kailath, T. (1980). *Linear systems*, volume 156. Prentice-Hall Englewood Cliffs, NJ.

Karatzas, I. and Shreve, S. (1998.). *Brownian motion and stochastic calculus*. Springer.

Khalil, H. (2002). *Nonlinear Systems*. Prentice-Hall.

Levinson, N. and Redheffer, R. M. (1970). *Complex variables*. Holden-Day.

Lofberg, J. (2004). Yalmip: A toolbox for modeling and optimization in matlab. In *Computer Aided Control Systems Design, 2004 IEEE International Symposium on*, pages 284–289. IEEE.

Ogata, K. (1970). Modern control engineering. *Prentice-Hall Electrical Engineering Series, Englewood Cliffs: Prentice-Hall,— c1970*.

Ogata, K. et al. (1995). *Discrete-time control systems*, volume 2. Prentice Hall Englewood Cliffs, NJ.

Rudin, W. (1964). *Principles of mathematical analysis*. McGraw-Hill New York.

Strang, G. (1976). *Linear Algebra and its Applications*. Academic Press.

Toh, K.-C., Todd, M. J., and Tütüncü, R. H. (1999). Sdpt3—a matlab software package for semidefinite programming, version 1.3. *Optimization methods and software*, 11(1-4):545–581.