Chapter 2

Linear System Theory

In this course, we will be dealing primarily with *linear* systems, a special class of systems for which a great deal is known. During the first half of the twentieth century, linear systems were analyzed using *frequency domain* (e.g., Laplace and z-transform) based approaches in an effort to deal with issues such as noise and bandwidth issues in communication systems. While they provided a great deal of intuition and were sufficient to establish some fundamental results, frequency domain approaches presented various drawbacks when control scientists started studying more complicated systems (containing multiple inputs and outputs, nonlinearities, noise, and so forth). Starting in the 1950's (around the time of the space race), control engineers and scientists started turning to *state-space* models of control systems in order to address some of these issues. These time-domain approaches are able to effectively represent concepts such as the internal state of the system, and also present a method to introduce optimality conditions into the controller design procedure. We will be using the state-space (or "modern") approach to control almost exclusively in this course, and the purpose of this chapter is to review some of the essential concepts in this area.

2.1 Discrete-Time Signals

Given a field $\mathbb{F}$, a *signal* is a mapping from a set of numbers to $\mathbb{F}$; in other words, signals are simply functions of the set of numbers that they operate on. More specifically:

- A *discrete-time* signal f is a mapping from the set of integers $\mathbb{Z}$ to $\mathbb{F}$, and is denoted by $f[k]$ for $k \in \mathbb{Z}$. Each instant k is also known as a *time-step*.
- A *continuous-time* signal f is a mapping from the set of real numbers $\mathbb{R}$ to $\mathbb{F}$ and is denoted by $f(t)$ for $t \in \mathbb{R}$.

One can obtain discrete-time signals by *sampling* continuous-time signals. Specifically, suppose that we are interested in the value of the signal $f(t)$ at times $t = 0, T, 2T, \ldots$, for some positive constant T. These values form a sequence $f(kT)$ for $k \in \mathbb{N}$, and if we simply drop the constant T from the notation (for convenience), we obtain the discrete-time sequence $f[k]$ for $k \in \mathbb{N}$. Since much of modern control deals with sampled versions of signals (due to the reliance on digital processing of such signals), we will be primarily working with discrete-time signals in this course, and focusing on the case where $\mathbb{F} = \mathbb{C}$ (the field of complex numbers).

2.2 Linear Time-Invariant Systems

Consider the system from Figure 1.1, with an input signal u and an output signal y. The system is either discrete-time or continuous-time, depending on the types of the signals. In our discussion, we will focus on the discrete-time case, but the results and definitions transfer in a straightforward manner to continuous-time systems.

In order to analyze and control the system, we will be interested in how the outputs respond to the inputs. We will be particularly interested in systems that satisfy the following property.

Definition 2.1 (Principle of Superposition)**.** Suppose that the output of the system is $y_1[k]$ in response to input $u_1[k]$ and $y_2[k]$ in response to input $u_2[k]$. The Principle of Superposition holds if the output of the system in response to the input $\alpha u_1[k] + \beta u_2[k]$ is $\alpha y_1[k] + \beta y_2[k]$, where α and β are arbitrary real numbers. Note that this must hold for *any* inputs $u_1[k]$ and $u_2[k]$.

The system is said to be **linear** if the Principle of Superposition holds. The system is said to be **time-invariant** if the output of the system is $y[k - \kappa]$ when the input is $u[k - \kappa]$ (i.e., a time-shifted version of the input produces an equivalent time-shift in the output). These concepts are illustrated in Fig. 2.1.

2.3 Mathematical Models

Mathematical models for many systems can be derived either from first principles (e.g., using Newton's Laws of motion for mechanical systems and Kirchoff's voltage and current laws for electrical systems). As noted earlier, many physical systems are inherently continuous-time, and thus their models would involve systems of differential equations. However, by *sampling* the system, one can essentially use discrete-time models to analyze such systems. In this course, we will predominantly be working with systems that can be represented by a certain mathematical form, and not focus too much on the explicit system itself. In other words, we will be interested in studying general properties and

Figure 2.1: (a) The Principle of Superposition. (b) The Time-Invariance Property.

analysis methods that can be applied to a *variety* of systems. To introduce the general mathematical system model that we will be considering, however, it will first be useful to consider a few examples.

2.3.1 A Simple Model for a Car

Consider again the model of the car from Chapter 1, with speed $v(t)$ at any time t, and acceleration input $a(t)$. Along the lines of Example 1.1, suppose that we *sample* the velocity of the car every T seconds, and that the acceleration is held constant between sampling times. We can then write

$$v[k+1] = v[k] + Ta[k], \tag{2.1}$$

where $v[k]$ is shorthand for $v(kT), k \in \mathbb{N}$, and $a[k]$ is the acceleration that is applied at time $t = kT$. Now, suppose that we wish to also consider the amount of fuel in the car at any given time-step k: let $g[k]$ denote this amount. We can consider a very simple model for fuel consumption, where the fuel decreases linearly with the distance traveled. Let $d[k]$ denote the distance traveled by the car between sampling times kT and $(k+1)T$, and recall from basic physics that under constant acceleration $a[k]$ and initial velocity $v[k]$, this distance is given by

$$d[k] = Tv[k] + \frac{T^2}{2}a[k].$$

Thus, if we let δ denote some coefficient that indicates how the fuel decreases with distance, we can write

$$g[k+1] = g[k] - \delta d[k] = g[k] - \delta Tv[k] - \delta\frac{T^2}{2}a[k]. \tag{2.2}$$

Equations (2.1) and (2.2) together form a two-state model of a car. We can put them together concisely using matrix-vector notation as follows:

$$\underbrace{\begin{bmatrix} v[k+1] \\ g[k+1] \end{bmatrix}}_{\mathbf{x}[k+1]} = \begin{bmatrix} 1 & 0 \\ -\delta T & 1 \end{bmatrix} \underbrace{\begin{bmatrix} v[k] \\ g[k] \end{bmatrix}}_{\mathbf{x}[k]} + \begin{bmatrix} T \\ -\delta \frac{T^2}{2} \end{bmatrix} a[k] \ . \tag{2.3}$$

The state of this system is given by the vector $\mathbf{x}[k] = \begin{bmatrix} v[k] & g[k] \end{bmatrix}'$, and the input is the acceleration $a[k]$, as before. If we consider the speedometer that provides a measurement of the speed at every sampling instant, the output of the system would be

$$s[k] = v[k] = \begin{bmatrix} 1 & 0 \end{bmatrix} \mathbf{x}[k].$$

We could also have a fuel sensor that measures $g[k]$ at each time-step; in this case, we would have two outputs, given by the vector

$$\mathbf{y}[k] = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \mathbf{x}[k],$$

where the first component corresponds to the sensor measurement of the speed, and the second component corresponds to the fuel sensor.

2.3.2 Longitudinal Dynamics of an F-8 Aircraft

Consider a model for the (sampled) linearized longitudinal dynamics of an F-8 aircraft [87]:

$$\underbrace{\begin{bmatrix} V[k+1] \\ \gamma[k+1] \\ \alpha[k+1] \\ q[k+1] \end{bmatrix}}_{\mathbf{x}[k+1]} = \begin{bmatrix} 0.9987 & -3.2178 & -4.4793 & -0.2220 \\ 0 & 1 & 0.1126 & 0.0057 \\ 0 & 0 & 0.8454 & 0.0897 \\ 0.0001 & -0.0001 & -0.8080 & 0.8942 \end{bmatrix} \underbrace{\begin{bmatrix} V[k] \\ \gamma[k] \\ \alpha[k] \\ q[k] \end{bmatrix}}_{\mathbf{x}[k]} + \begin{bmatrix} -0.0329 \\ 0.0131 \\ -0.0137 \\ -0.0092 \end{bmatrix} \mathbf{u}[k], \tag{2.4}$$

where $V[k]$ is the velocity of the aircraft, $\gamma[k]$ is the flight-path angle, $\alpha[k]$ is the angle-of-attack, and $q[k]$ is the pitch rate. The input to the aircraft is taken to be the deflection of the elevator flaps.

The output of the system will depend on the sensors that are installed on the aircraft. For example, if there is a sensor to measure the velocity and the pitch rate, the output would be given by

$$\mathbf{y}[k] = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \mathbf{x}[k].$$

2.3.3 Linear Feedback Shift Register

Consider, a *linear feedback shift register (LFSR)*, which is a digital circuit used to implement function such as random number generators and "noise" sequences in computers.

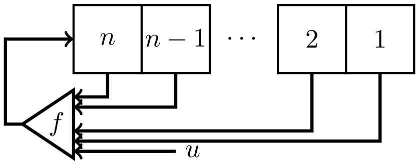

Figure 2.2: A Linear Feedback Shift Register

The LFSR consists of n registers that are chained together. At each clock-cycle k, each register takes the value of the next register in the chain. The last register's value is computed as a function of the values of other registers further up the chain, and perhaps with an additional input $u[k]$. If we denote the value stored in register i at time-step (or clock-cycle) k by $x_i[k]$, we obtain the model

$$\begin{aligned} x_1[k+1] &= x_2[k] \\ x_2[k+1] &= x_3[k] \\ &\vdots \\ x_{n-1}[k+1] &= x_n[k] \\ x_n[k+1] &= \alpha_0 x_1[k] + \alpha_1 x_2[k] + \cdots + \alpha_{n-1} x_n[k] + \beta u[k], \end{aligned} \tag{2.5}$$

for some scalars $\alpha_0, \alpha_1, \ldots, \alpha_{n-1}, \beta$. These scalars and input are usually chosen so that the sequence of values generated by register 1 exhibits certain behavior (e.g., simulates the generation of a "random" sequence of values). Let $y[k] = x_1[k]$ denote the output of the system. One can write the above equations more compactly by defining the *state vector*

$$\mathbf{x}[k] \triangleq \begin{bmatrix} x_1[k] & x_2[k] & \cdots & x_n[k] \end{bmatrix}' ,$$

from which we obtain

$$\begin{aligned} \mathbf{x}[k+1] &= \underbrace{\begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \\ \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & \cdots & \alpha_{n-1} \end{bmatrix}}_{\mathbf{A}} \mathbf{x}[k] + \underbrace{\begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ \beta \end{bmatrix}}_{\mathbf{B}} u[k] \\ y[k] &= \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \end{bmatrix}}_{\mathbf{C}} \mathbf{x}[k]. \end{aligned}$$

2.4 State-Space Models

In the last section, we saw some examples of physical systems that can be modeled via a set of discrete-time equations, which were then put into a matrix-vector form. Such forms known as *state-space models* of linear systems, and can generally have multiple inputs and outputs to the system, with general system matrices.[1]

The state-space model for a discrete-time linear system is given by

$$\begin{aligned} \mathbf{x}[k+1] &= \mathbf{A}\mathbf{x}[k] + \mathbf{B}\mathbf{u}[k] \\ \mathbf{y}[k] &= \mathbf{C}\mathbf{x}[k] + \mathbf{D}\mathbf{u}[k] \ . \end{aligned} \tag{2.6}$$

The state-space model of a continuous-time linear system is given by

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} \\ \mathbf{y} &= \mathbf{C}\mathbf{x} + \mathbf{D}\mathbf{u} \ . \end{aligned} \tag{2.7}$$

- The vector $\mathbf{x}$ is called the **state vector** of the system. We will denote the number of states in the system by n, so that $\mathbf{x} \in \mathbb{R}^n$. The quantity n is often called the **order** or **dimension** of the system.
- In general, we might have multiple inputs $u_1, u_2, \ldots, u_m$ to the system. In this case, we can define an **input vector** $\mathbf{u} = \begin{bmatrix} u_1 & u_2 & \cdots & u_m \end{bmatrix}'$.
- In general, we might have multiple outputs $y_1, y_2, \ldots, y_p$. In this case, we can define the **output vector** $\mathbf{y} = \begin{bmatrix} y_1 & y_2 \cdots & y_p \end{bmatrix}'$. Note that each of these outputs represents a sensor measurement of some of the states of the system.
- The **system matrix A** is an $n \times n$ matrix representing how the states of the system affect each other.
- The **input matrix B** is an $n \times m$ matrix representing how the inputs to the system affect the states.
- The **output matrix C** is a $p \times n$ matrix representing the portions of the states that are measured by the outputs.
- The **feedthrough matrix D** is a $p \times m$ matrix representing how the inputs affect the outputs directly (i.e., without going through the states first).

[1] Although we will be focusing on linear systems, many practical systems are *nonlinear*. Since state-space models are time-domain representations of systems, they can readily capture nonlinear dynamics. When the states of the system stay close to some nominal operating point, nonlinear systems can often be *linearized*, bringing them into the form (2.6) or (2.7). We will not discuss nonlinear systems in too much further detail in this course.

2.4.1 Transfer Functions of Linear State-Space Models

While the state-space models (2.6) and (2.7) are a time-domain representation of systems, one can also convert them to the frequency domain by taking the z-transform (or Laplace transform in continuous-time). Specifically, if we take the z-transform of (2.6), we obtain:

$$\begin{aligned} z\mathbf{X}(z) - z\mathbf{x}(0) &= \mathbf{A}\mathbf{X}(z) + \mathbf{B}\mathbf{U}(z) \\ \mathbf{Y}(z) &= \mathbf{C}\mathbf{X}(z) + \mathbf{D}\mathbf{U}(z) \ . \end{aligned}$$

Note that this includes the initial conditions of all the states. The first equation can be rearranged to solve for $\mathbf{X}(z)$ as follows:

$$(z\mathbf{I} - \mathbf{A})\mathbf{X}(z) = z\mathbf{x}(0) + \mathbf{B}\mathbf{U}(z) \Leftrightarrow \mathbf{X}(z) = (z\mathbf{I} - \mathbf{A})^{-1}z\mathbf{x}(0) + (z\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}\mathbf{U}(z) \ .$$

Substituting this into the equation for $\mathbf{Y}(z)$, we obtain

$$\mathbf{Y}(z) = \mathbf{C}(z\mathbf{I} - \mathbf{A})^{-1}z\mathbf{x}(0) + \big(\mathbf{C}(z\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D}\big)\,\mathbf{U}(z) \ .$$

The transfer function of the state-space model $\mathbf{x}[k+1] = \mathbf{A}\mathbf{x}[k] + \mathbf{B}\mathbf{u}[k]$, $\mathbf{y}[k] = \mathbf{C}\mathbf{x}[k] + \mathbf{D}\mathbf{u}[k]$ (when $\mathbf{x}(0) = 0$) is

$$\mathbf{H}(z) = \mathbf{C}(z\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D} \ . \tag{2.8}$$

Note that $\mathbf{H}(z)$ is a $p \times m$ matrix, and thus it is a generalization of the transfer function for standard single-input single-output systems. In fact, it is a matrix where entry i, j is a transfer function describing how the j–th input affects the i–th output.

Example 2.1. Calculate the transfer function for the state space model

$$\mathbf{x}[k+1] = \underbrace{\begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix}}_{\mathbf{A}} \mathbf{x}[k] + \underbrace{\begin{bmatrix} 0 \\ 4 \end{bmatrix}}_{\mathbf{B}} u[k], \quad \mathbf{y}[k] = \underbrace{\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}}_{\mathbf{C}} \mathbf{x}[k] + \underbrace{\begin{bmatrix} 3 \\ 0 \end{bmatrix}}_{\mathbf{D}} u[k] \ .$$

$$\begin{aligned} \mathbf{H}(z) &= \mathbf{C}\left(z\mathbf{I} - \mathbf{A}\right)^{-1}\mathbf{B} + \mathbf{D} \\ &= \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} z & -1 \\ 2 & z+3 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 3 \\ 0 \end{bmatrix} \\ &= \frac{1}{z^2 + 3z + 2} \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} z+3 & 1 \\ -2 & z \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 3 \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} \frac{3z^2+10z+9}{z^2+3z+2} \\ \frac{1}{z+2} \end{bmatrix} . \end{aligned}$$

2.4.2 State-Space Realizations and Similarity Transformations

Suppose we have a linear system with transfer function $\mathbf{H}(z)$ (which can be a matrix, in general). We have seen that the transfer function is related to the matrices in the state space model via (2.8). Recall that the transfer function describes how the input to the system affects the output (when the initial state of the system is zero). In some sense, this might seem to indicate that the exact representation of the internal states of the system might be irrelevant, as long as the input-output behavior is preserved. In this section, we will see that there are multiple *state-space realizations* for a given system that correspond to the same transfer function.

Consider any particular state-space model of the form (2.6). Now, let us choose an arbitrary invertible $n \times n$ matrix $\mathbf{T}$, and define a new state vector

$$\bar{\mathbf{x}}[k] = \mathbf{T}\mathbf{x}[k] \ .$$

In other words, the states in the vector $\bar{\mathbf{x}}[k]$ are linear combinations of the states in the vector $\mathbf{x}[k]$. Since $\mathbf{T}$ is a constant matrix, we have

$$\begin{aligned}\bar{\mathbf{x}}[k+1] = \mathbf{T}\mathbf{x}[k+1] = \mathbf{T}\mathbf{A}\mathbf{x}[k] + \mathbf{T}\mathbf{B}\mathbf{u}[k] &= \underbrace{\mathbf{T}\mathbf{A}\mathbf{T}^{-1}}_{\bar{\mathbf{A}}}\bar{\mathbf{x}}[k] + \underbrace{\mathbf{T}\mathbf{B}}_{\bar{\mathbf{B}}}\mathbf{u}[k] \\ \mathbf{y} = \mathbf{C}\mathbf{x}[k] + \mathbf{D}\mathbf{u}[k] &= \underbrace{\mathbf{C}\mathbf{T}^{-1}}_{\bar{\mathbf{C}}}\bar{\mathbf{x}}[k] + \mathbf{D}\mathbf{u}[k] \ .\end{aligned}$$

Thus, after this transformation, we obtain the new state-space model

$$\begin{aligned}\bar{\mathbf{x}}[k+1] &= \bar{\mathbf{A}}\bar{\mathbf{x}}[k] + \bar{\mathbf{B}}\mathbf{u}[k] \\ \mathbf{y} &= \bar{\mathbf{C}}\bar{\mathbf{x}}[k] + \mathbf{D}\mathbf{u}[k] \ .\end{aligned}$$

Note that the inputs and outputs were *not* affected by this transformation; only the internal state vector and matrices changed. The transfer function corresponding to this model is given by

$$\begin{aligned}\bar{\mathbf{H}}(z) = \bar{\mathbf{C}}(z\mathbf{I} - \bar{\mathbf{A}})^{-1}\bar{\mathbf{B}} + \mathbf{D} &= \mathbf{C}\mathbf{T}^{-1}(z\mathbf{I} - \mathbf{T}\mathbf{A}\mathbf{T}^{-1})^{-1}\mathbf{T}\mathbf{B} + \mathbf{D} \\ &= \mathbf{C}\mathbf{T}^{-1}(z\mathbf{T}\mathbf{T}^{\mathbf{-1}} - \mathbf{T}\mathbf{A}\mathbf{T}^{-1})^{-1}\mathbf{T}\mathbf{B} + \mathbf{D} \\ &= \mathbf{C}\mathbf{T}^{-1}\mathbf{T}(z\mathbf{I} - \mathbf{A})^{-1}\mathbf{T}^{-1}\mathbf{T}\mathbf{B} + \mathbf{D} \\ &= \mathbf{C}(z\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D} \\ &= \mathbf{H}(z) \ .\end{aligned}$$

Thus the transfer function for the realization with state-vector $\bar{\mathbf{x}}$ is the same as the transfer function for the realization with state-vector $\mathbf{x}$. For this reason, the transformation $\bar{\mathbf{x}} = \mathbf{T}\mathbf{x}$ is called a **similarity transformation**. Since $\mathbf{T}$ can be *any* invertible matrix, and since there are an infinite number of invertible $n \times n$ matrices to choose from, we see that there are an **infinite number of realizations** for any given transfer function $\mathbf{H}(z)$.

Similarity transformations are a very useful tool to analyze the behavior of linear systems, as we will see in later sections.

2.5 Stability of Linear Systems

Consider the system

$$\mathbf{x}[k+1] = \mathbf{A}\mathbf{x}[k] \ , \tag{2.9}$$

without any inputs. This is known as an *autonomous* system. The following definition plays a central role in control theory.

Definition 2.2 (Stability)**.** The linear system $\mathbf{x}[k+1] = \mathbf{A}\mathbf{x}[k]$ is said to be *stable* if

$$\lim_{k\to\infty} \mathbf{x}[k] = \mathbf{0}$$

starting from any initial state $\mathbf{x}[0]$.

To obtain conditions for stability, it is first instructive to consider the scalar system $x[k+1] = \alpha x[k]$, where $\alpha \in \mathbb{R}$. Since $x[1] = \alpha x[0]$, $x[2] = \alpha x[1] = \alpha^2 x[0]$, and so forth, we have $x[k] = \alpha^k x[0]$. Now, in order for $x[k]$ to go to zero regardless of the value of $x[0]$, we must have $\alpha^k \to 0$ as $k \to \infty$, and this happens if and only if $|\alpha| < 1$. This is the necessary and sufficient condition for stability of a scalar linear system.

One can extend this to general state-space models of the form (2.9). To give the main idea of the proof, suppose that $\mathbf{A}$ is diagonalizable and write

$$\mathbf{A} = \mathbf{T}\Lambda\mathbf{T}^{-1} = \mathbf{T}\begin{bmatrix} \lambda_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \lambda_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \lambda_n \end{bmatrix}\mathbf{T}^{-1},$$

where each λ_i is an eigenvalue of $\mathbf{A}$. It is easy to verify that

$$\mathbf{x}[k] = \mathbf{A}^k\mathbf{x}[0] = \mathbf{T}\Lambda^k\mathbf{T}^{-1}\mathbf{x}[0] = \mathbf{T}\begin{bmatrix} \lambda_1^k & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \lambda_2^k & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \lambda_n^k \end{bmatrix}\mathbf{T}^{-1}\mathbf{x}[0].$$

This expression goes to zero for any $\mathbf{x}[0]$ if and only if $|\lambda_i| < 1$ for all $i \in \{1, 2, \ldots, n\}$. The proof for the most general case (where $\mathbf{A}$ is not diagonalizable) can be obtained by considering the Jordan form of $\mathbf{A}$ (see Appendix A.4.5); we will omit the mathematical details because they do not add to much to the discussion or understanding here. Thus, we obtain the following fundamental result.

Theorem 2.1. *The linear system* $\mathbf{x}[k+1] = \mathbf{A}\mathbf{x}[k]$ *is stable if and only if all eigenvalues of* $\mathbf{A}$ *have magnitude smaller than* 1 *(i.e., they are contained within the open unit circle in the complex plane).*

Frequently, the system under consideration is not stable (i.e., the $\mathbf{A}$ matrix contains eigenvalues of magnitude larger than 1), and the objective is to choose the inputs to the system so that $\mathbf{x}[k] \to 0$ as $k \to \infty$. We will study conditions under which this is possible in the next few sections.

2.6 Properties of Linear Systems

We now turn our attention to analyzing the state-space model (2.6) for the purpose of controlling the system. There are several properties of such systems that we will be studying.

2.6.1 Controllability

Definition 2.3 (Controllability)**.** The system (2.6) is said to be *controllable* if, for any initial state $\mathbf{x}[0]$ and any desired state $\mathbf{x}^*$, there is a nonnegative integer L and a sequence of inputs $\mathbf{u}[0], \mathbf{u}[1], \ldots, \mathbf{u}[L]$ such that $\mathbf{x}[L+1] = \mathbf{x}^*$.

To derive conditions on the system matrices $\mathbf{A}$ and $\mathbf{B}$ under which the system is controllable, suppose we start at some state $\mathbf{x}[0]$ at time-step 0, and note that:

$$\begin{aligned}
\mathbf{x}[1] &= \mathbf{A}\mathbf{x}[0] + \mathbf{B}\mathbf{u}[0]\\
\mathbf{x}[2] &= \mathbf{A}\mathbf{x}[1] + \mathbf{B}\mathbf{u}[1] = \mathbf{A}^2\mathbf{x}[0] + \mathbf{A}\mathbf{B}\mathbf{u}[0] + \mathbf{B}\mathbf{u}[1]\\
\mathbf{x}[3] &= \mathbf{A}\mathbf{x}[2] + \mathbf{B}\mathbf{u}[2] = \mathbf{A}^3\mathbf{x}[0] + \mathbf{A}^2\mathbf{B}\mathbf{u}[0] + \mathbf{A}\mathbf{B}\mathbf{u}[1] + \mathbf{B}\mathbf{u}[2].
\end{aligned}$$

Continuing in this way, we can write

$$\mathbf{x}[L+1] - \mathbf{A}^{L+1}\mathbf{x}[0] = \underbrace{\begin{bmatrix}\mathbf{A}^L\mathbf{B} & \mathbf{A}^{L-1}\mathbf{B} & \cdots & \mathbf{B}\end{bmatrix}}_{\mathcal{C}_L} \underbrace{\begin{bmatrix}\mathbf{u}[0]\\ \mathbf{u}[1]\\ \vdots \\ \mathbf{u}[L]\end{bmatrix}}_{\mathbf{u}[0:L]} .$$

The matrix $\mathcal{C}_L$ is called the *controllability matrix* for the pair $(\mathbf{A}, \mathbf{B})$. In order to go from $\mathbf{x}[0]$ to any value $\mathbf{x}[L+1]$, it must be the case that

$$\mathbf{x}[L+1] - \mathbf{A}^{L+1}\mathbf{x}[0] \in \mathcal{R}\left(\mathcal{C}_L\right),$$

where $\mathcal{R}(\cdot)$ denotes the range space of a matrix (see Appendix A.4.1). If we want to be able to go from any arbitrary initial state to any other arbitrary final state in $L+1$ time-steps, it must be the case that $\mathcal{R}\left(\mathcal{C}_L\right) = \mathbb{R}^n$, which is equivalent to saying that $\text{rank}(\mathcal{C}_L) = n$. If this condition is satisfied, then we can find n linearly independent columns within $\mathcal{C}_L$, and select the inputs $\mathbf{u}[0], \mathbf{u}[1], \ldots, \mathbf{u}[L]$ to combine those columns in such a way that we can obtain any $\mathbf{x}[L+1]$. However, if the rank of $\mathcal{C}_L$ is less than n, then there might be some $\mathbf{x}[L+1]$ that we cannot obtain. In this case, we can wait a few more time-steps and hope that the rank of the matrix $\mathcal{C}_L$ increases to n. How long should we wait?

To answer this question, note that the rank of $\mathcal{C}_L$ is a nondecreasing function of L, and bounded above by n. Suppose ν is the first integer for which $\text{rank}(\mathcal{C}_\nu) = \text{rank}(\mathcal{C}_{\nu-1})$. This is equivalent to saying that the extra columns in $\mathcal{C}_\nu$ (given by $\mathbf{A}^\nu\mathbf{B}$) are all linearly dependent on the columns in $\mathcal{C}_{\nu-1}$; mathematically, there exists a matrix $\mathbf{K}$ such that

$$\mathbf{A}^\nu\mathbf{B} = \begin{bmatrix}\mathbf{A}^{\nu-1}\mathbf{B} & \mathbf{A}^{\nu-2}\mathbf{B} & \cdots & \mathbf{B}\end{bmatrix}\mathbf{K} \ .$$

In turn, this implies that

$$\mathbf{A}^{\nu+1}\mathbf{B} = \mathbf{A}\mathbf{A}^\nu\mathbf{B} = \mathbf{A}\mathcal{C}_{\nu-1}\mathbf{K} = \begin{bmatrix}\mathbf{A}^\nu\mathbf{B} & \mathbf{A}^{\nu-1}\mathbf{B} & \cdots & \mathbf{A}\mathbf{B}\end{bmatrix}\mathbf{K} \ ,$$

and so the matrix $\mathbf{A}^{\nu+1}\mathbf{B}$ can be written as a linear combination of the columns in $\mathcal{C}_\nu$ (which can themselves be written as linear combinations of columns in $\mathcal{C}_{\nu-1}$). Continuing in this way, we see that

$$\text{rank}(\mathcal{C}_0) < \text{rank}(\mathcal{C}_1) < \cdots < \text{rank}(\mathcal{C}_{\nu-1}) = \text{rank}(\mathcal{C}_\nu) = \text{rank}(\mathcal{C}_{\nu+1}) = \cdots \ ,$$

i.e., the rank of $\mathcal{C}_L$ monotonically increases with L until $L = \nu - 1$, at which point it stops increasing. Since the matrix $\mathbf{B}$ contributes $\text{rank}(\mathbf{B})$ linearly independent columns to the controllability matrix, the rank of the controllability matrix can increase for at most $n - \text{rank}(\mathbf{B})$ time-steps before it reaches its maximum value, and so the integer ν is upper bounded as $\nu \le n - \text{rank}(\mathbf{B}) + 1$. This yields the following result.

Theorem 2.2. *Consider the system* (2.6)*, where* $\mathbf{x}[k] \in \mathbb{R}^n$*. For any positive integer* L*, define the* **controllability matrix**

$$\mathcal{C}_L = \begin{bmatrix}\mathbf{A}^L\mathbf{B} & \mathbf{A}^{L-1}\mathbf{B} & \cdots & \mathbf{B}\end{bmatrix} \ . \tag{2.10}$$

The system is controllable if and only if $\mathit{rank}(\mathcal{C}_{n-\mathit{rank}(\mathbf{B})}) = n$.

In the linear systems literature, the integer ν is called the *controllability index* of the pair $(\mathbf{A}, \mathbf{B})$. For simplicity, one often uses the fact that $n - \text{rank}(\mathbf{B}) \leq n - 1$, and just checks the rank of $\mathcal{C}_{n-1}$ to verify controllability.

Example 2.2. Consider the system given by

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

The controllability matrix for this system is

$$\mathcal{C}_{n-\text{rank}(\mathbf{B})} = \mathcal{C}_1 = \begin{bmatrix} \mathbf{AB} & \mathbf{B} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -3 & 1 \end{bmatrix}.$$

This matrix has rank equal to 2, and so the system is controllable. Specifically, since

$$\mathbf{x}[2] = \mathbf{A}^2\mathbf{x}[0] + \mathcal{C}_1 \begin{bmatrix} u[0] \\ u[1] \end{bmatrix},$$

we can go from any initial state $\mathbf{x}[0]$ to any state $\mathbf{x}[2]$ at time-step 2 simply by applying the inputs

$$\begin{bmatrix} u[0] \\ u[1] \end{bmatrix} = \mathcal{C}_1^{-1} \left(\mathbf{x}[2] - \mathbf{A}^2\mathbf{x}[0] \right).$$

Example 2.3. Consider the system given by

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & -2 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

The controllability matrix for this system is

$$\mathcal{C}_{n-\text{rank}(\mathbf{B})} = \mathcal{C}_1 = \begin{bmatrix} \mathbf{AB} & \mathbf{B} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}.$$

This matrix only has rank 1, and thus the system is not controllable. Specifically, starting from an initial state of zero, one can never drive the second state to a nonzero value.

2.6.2 Observability

Definition 2.4 (Observability)**.** The system is said to be *observable* if, for any initial state $\mathbf{x}[0]$, and for any *known* sequence of inputs $\mathbf{u}[0], \mathbf{u}[1], \ldots$, there is a positive integer L such that $\mathbf{x}[0]$ can be recovered from the outputs $\mathbf{y}[0], \mathbf{y}[1], \ldots, \mathbf{y}[L]$.

To relate the concept of observability to the system matrices, if we simply iterate the output equation in (2.6) for $L+1$ time-steps, we get:

$$\underbrace{\begin{bmatrix}\mathbf{y}[0]\\ \mathbf{y}[1]\\ \mathbf{y}[2]\\ \vdots\\ \mathbf{y}[L]\end{bmatrix}}_{\mathbf{y}[0:L]} = \underbrace{\begin{bmatrix}\mathbf{C}\\ \mathbf{CA}\\ \mathbf{CA}^2\\ \vdots\\ \mathbf{CA}^L\end{bmatrix}}_{\mathcal{O}_L}\mathbf{x}[0] + \underbrace{\begin{bmatrix}\mathbf{D} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{CB} & \mathbf{D} & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{CAB} & \mathbf{CB} & \mathbf{D} & \cdots & \mathbf{0}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ \mathbf{CA}^{L-1}\mathbf{B} & \mathbf{CA}^{L-2}\mathbf{B} & \mathbf{CA}^{L-3}\mathbf{B} & \cdots & \mathbf{D}\end{bmatrix}}_{\mathcal{J}_L}\underbrace{\begin{bmatrix}\mathbf{u}[0]\\ \mathbf{u}[1]\\ \mathbf{u}[2]\\ \vdots\\ \mathbf{u}[L]\end{bmatrix}}_{\mathbf{u}[0:L]}. \quad (2.11)$$

The matrix $\mathcal{O}_L$ is called the *observability* matrix for the pair $(\mathbf{A},\mathbf{C})$, and the matrix $\mathcal{J}_L$ is called the *invertibility matrix* for the tuple $(\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D})$; this terminology will become clear in next section. Rearranging the above equation, we obtain

$$\mathbf{y}[0:L] - \mathcal{J}_L\mathbf{u}[0:L] = \mathcal{O}_L\mathbf{x}[0].$$

Since the inputs to the system are assumed to be *known* in this case, the entire left hand side of the above equation is known. Thus, the objective is to uniquely recover $\mathbf{x}[0]$ from the above equation; this is possible if and only if $\text{rank}(\mathcal{O}_L) = n$. In this case, the system is said to be *observable*. As in the case of the controllability matrix, one can show that there exists an integer μ such that

$$\text{rank}(\mathcal{O}_0) < \text{rank}(\mathcal{O}_1) < \cdots < \text{rank}(\mathcal{O}_{\mu-1}) = \text{rank}(\mathcal{O}_\mu) = \text{rank}(\mathcal{O}_{\mu+1}) = \cdots\ ,$$

i.e., the rank of $\mathcal{O}_L$ monotonically increases with L until $L = \mu - 1$, at which point it stops increasing. Since the matrix $\mathbf{C}$ contributes $\text{rank}(\mathbf{C})$ linearly independent rows to the controllability matrix, the rank of the observability matrix can increase for at most $n - \text{rank}(\mathbf{C})$ time-steps before it reaches its maximum value, and so the integer μ is upper bounded as $\mu \le n - \text{rank}(\mathbf{C}) + 1$. This yields the following result.

Theorem 2.3. *Consider the system* (2.6)*, where* $\mathbf{x}[k] \in \mathbb{R}^n$*. For any positive integer* L*, define the* **observability matrix**

$$\mathcal{O}_L = \begin{bmatrix}\mathbf{C}\\ \mathbf{CA}\\ \mathbf{CA}^2\\ \vdots\\ \mathbf{CA}^L\end{bmatrix}. \quad (2.12)$$

The system is observable if and only if $\text{rank}(\mathcal{O}_{n-\text{rank}(\mathbf{C})}) = n$.

The integer μ is called the *observability index* of the system.

Remark 2.1. It is easy to show that the pair $(\mathbf{A}, \mathbf{C})$ is observable if and only if the pair $(\mathbf{A}', \mathbf{C}')$ is controllable; simply transpose the observability matrix and rearrange the columns (which does not change the rank of the observability matrix) to resemble the controllability matrix. Thus, controllability and observability are known as *dual* properties of linear systems. Note that this does *not* mean that a given system that is controllable is also observable (since the controllability matrix involves the matrix $\mathbf{B}$, and the observability matrix involves $\mathbf{C}$).

Example 2.4. Consider the pair $\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix}$, $\mathbf{C} = \begin{bmatrix} 1 & 0 \end{bmatrix} \mathbf{x}$, with no inputs to the system. The observability matrix for this pair is

$$\mathcal{O}_{n-\text{rank}(\mathbf{C})} = \mathcal{O}_1 = \begin{bmatrix} \mathbf{C} \\ \mathbf{CA} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix},$$

which has rank 2, and thus the pair is observable. The initial state of the system can be recovered as follows:

$$\mathbf{y}[0:1] = \mathcal{O}_1 \mathbf{x}[0] \quad \Rightarrow \mathcal{O}_1^{-1} \mathbf{y}[0:1] = \mathbf{x}[0].$$

2.6.3 Invertibility

In the last section, we assumed that the inputs to the system were completely known; this allowed us to subtract them out from the outputs of the system, and then recover the initial state (provided that the system was observable). However, there may be cases where some or all of the inputs to the system are completely unknown and arbitrary, the system is called a *linear system with unknown inputs* [35]. For such systems, it is often of interest to "invert" the system in order to reconstruct some or all of the unknown inputs (assuming that the initial state is known), and this problem has been studied under the moniker of *dynamic system inversion* [72, 75]. This concept will be very useful when we discuss the diagnosis of faults and attacks in linear systems, since such events can often be modeled via unknown inputs to the system.

Definition 2.5 (Invertibility)**.** The system (2.6) is said to have an L-delay inverse if it is possible to uniquely recover the input $\mathbf{u}[k]$ from the outputs of the system up to time-step $\mathbf{y}[k+L]$ (for some nonnegative integer L), assuming that the initial state $\mathbf{x}[0]$ is known. The system is *invertible* if it has an L-delay inverse for some finite L. The least integer L for which an L-delay inverse exists is called the inherent delay of the system.

To illustrate the idea, let us start by considering an example.

Example 2.5. Consider the system

$$\mathbf{x}[k+1] = \begin{bmatrix} 0 & 1 \\ 2 & -3 \end{bmatrix} \mathbf{x}[k] + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \mathbf{u}[k], \ \ \mathbf{y}[k] = \begin{bmatrix} 1 & 0 \end{bmatrix} \mathbf{x}[k].$$

Clearly $\mathbf{y}[k]$ provides no information about $\mathbf{u}[k]$. Similarly, $\mathbf{y}[k+1] = \mathbf{CAx}[k] + \mathbf{CBu}[k] = \begin{bmatrix} 0 & 1 \end{bmatrix} \mathbf{x}[k]$, which again does not contain any information about $\mathbf{u}[k]$. However, $\mathbf{y}[k+2] = \mathbf{CA}^2\mathbf{x}[k] + \mathbf{CABu}[k] + \mathbf{CBu}[k+1] = \begin{bmatrix} 2 & -3 \end{bmatrix} \mathbf{x}[k] + \mathbf{u}[k]$, and thus we can recover $\mathbf{u}[k]$ as $\mathbf{y}[k+2] - \begin{bmatrix} 2 & -3 \end{bmatrix} \mathbf{x}[k]$, provided that we knew $\mathbf{x}[k]$. Specifically, if we knew $\mathbf{x}[0]$, we would be able to recover $\mathbf{u}[0]$ from the above expression, and then we could determine $\mathbf{x}[1] = \mathbf{Ax}[0] + \mathbf{Bu}[0]$. We could then repeat the procedure to find $\mathbf{u}[k]$ (and $\mathbf{x}[k]$) for all $k \in \mathbb{N}$).

To come up with a systematic procedure to analyze invertibility of systems, consider again the output of the linear system (2.6) over $L+1$ time-steps for any nonnegative integer L; rearranging (2.11), we see that

$$\mathbf{y}[0:L] - \mathcal{O}_L\mathbf{x}[0] = \mathcal{J}_L\mathbf{u}[0:L], \tag{2.13}$$

where the left side is now assumed to be completely known. The matrix $\mathcal{J}_L$ will completely characterize our ability to recover the inputs to the system. First, note from (2.11) that the last Lm columns of $\mathcal{J}_L$ have the form $\begin{bmatrix} \mathbf{0} \\ \mathcal{J}_{L-1} \end{bmatrix}$. The rank of $\mathcal{J}_L$ is thus equal to the number of linearly independent columns from the last Lm columns (given by rank$(\mathcal{J}_{L-1})$, plus any additional linearly independent columns from the first m columns. Thus,

$$\text{rank}(\mathcal{J}_L) \le m + \text{rank}\left(\begin{bmatrix} \mathbf{0} \\ \mathcal{J}_{L-1} \end{bmatrix}\right) = m + \text{rank}(\mathcal{J}_{L-1}), \tag{2.14}$$

for all nonnegative integers L, where we define rank$(\mathcal{J}_{-1}) = 0$ for convenience.

Now, note that the input $\mathbf{u}[0]$ enters equation (2.13) through the first m columns of the matrix $\mathcal{J}_L$. Thus, in order to recover $\mathbf{u}[0]$, it must be the case that:

1. The first m columns of $\mathcal{J}_L$ are linearly independent of each other (otherwise there exists some nonzero $\mathbf{u}[0]$ such that the first m columns times $\mathbf{u}[0]$ is the zero vector, which is indistinguishable from the case where $\mathbf{u}[0] = \mathbf{0}$).
2. The first m columns of $\mathcal{J}_L$ are linearly independent of *all other* columns of $\mathcal{J}_L$ (otherwise, there exists some nonzero $\mathbf{u}[0]$ and some nonzero $\mathbf{u}[1:L]$ such that $\mathcal{J}_L\mathbf{u}[0:L] = \mathbf{0}$, which is indistinguishable from case where $\mathbf{u}[0:L] = \mathbf{0}$).

If both of the above conditions are satisfied, then one can find a matrix $\mathbf{P}$ such that $\mathbf{P}\mathcal{J}_L = \begin{bmatrix} \mathbf{I}_m & \mathbf{0} \end{bmatrix}$, which means that the input $\mathbf{u}[0]$ can be recovered as

$$\mathbf{P}\left(\mathbf{y}[0:L] - \mathcal{O}_L\mathbf{x}[0]\right) = \mathbf{P}\mathcal{J}_L\mathbf{u}[0:L] = \mathbf{u}[0] \ .$$

Since $\mathbf{u}[0]$ is now known, one can obtain $\mathbf{x}[1] = \mathbf{Ax}[0] + \mathbf{Bu}[0]$, and can repeat the process to obtain $\mathbf{u}[k]$ for all positive integers k.

The condition that the first m columns of $\mathcal{J}_L$ be linearly independent of all other columns and of each other is equivalent to saying that

$$\text{rank}(\mathcal{J}_L) = m + \text{rank}\left(\begin{bmatrix} \mathbf{0} \\ \mathcal{J}_{L-1} \end{bmatrix}\right) = m + \text{rank}(\mathcal{J}_{L-1}),$$

i.e., equality holds in (2.14). Thus, to check for invertibility of the linear system (2.6), we can start with $\mathcal{J}_0 = \mathbf{D}$, and increase L until we find $\text{rank}(\mathcal{J}_L) = m + \text{rank}(\mathcal{J}_{L-1})$. At what point should we stop increasing L and announce that the system is *not* invertible? To answer this question, we will use the following argument from [72]. First, suppose that the system is not invertible for $L = 0, 1, \ldots, n$. Then from (2.14), we have

$$\text{rank}(\mathcal{J}_n) \leq m - 1 + \text{rank}(\mathcal{J}_{n-1)} \leq 2(m-1) + \text{rank}(\mathcal{J}_{n-2}) \leq \cdots \leq (n+1)(m-1).$$

Note that we use $m-1$ in each of the above inequalities because we know that (2.14) holds with strict inequality (due to the fact that the system is not invertible for those delays). Based on the above inequality, the null space of $\mathcal{J}_n$ has dimension

$$(n+1)m - \text{rank}(\mathcal{J}_n) \geq (n+1)m - (n+1)(m-1) = n+1.$$

Let $\mathbf{N}$ be a matrix whose columns from a basis for the null space of $\mathcal{J}_n$, and note that $\mathbf{N}$ has at least $n+1$ columns. Thus, any input of the form $\mathbf{u}[0:n] = \mathbf{N}\mathbf{v}$ for some vector $\mathbf{v}$ would produce $\mathcal{J}_n\mathbf{u}[0:n] = \mathcal{J}_n\mathbf{N}\mathbf{v} = \mathbf{0}$. Now, also note that

$$\mathbf{x}[n+1] = \mathbf{A}^{n+1}\mathbf{x}[0] + \mathcal{C}_n\mathbf{u}[0:n] = \mathbf{A}^{n+1}\mathbf{x}[0] + \mathcal{C}_n\mathbf{N}\mathbf{v}.$$

Note that the matrix $\mathcal{C}_n\mathbf{N}$ has n rows and at least $n+1$ columns; thus it has a null space of dimension at least one. Thus, if we pick the vector $\mathbf{v}$ to be any vector in this null space, we see that $\mathbf{x}[n+1] = \mathbf{A}^{n+1}\mathbf{x}[0]$ and $\mathbf{y}[0:n] = \mathcal{O}_n\mathbf{x}[0]$. In other words, the input sequence $\mathbf{u}[0:n] = \mathbf{N}\mathbf{v}$ chosen in this way produces the same output over $n+1$ time-steps as the input sequence $\mathbf{u}[0:n] = \mathbf{0}$, and also leaves the state $\mathbf{x}[n+1]$ in the same position as the all zero input. If the input is $\mathbf{u}[k] = \mathbf{0}$ for all $k \geq n+1$, we see that we can never determine whether $\mathbf{u}[0:n] = \mathbf{N}\mathbf{v}$ or $\mathbf{u}[0:n] = \mathbf{0}$. Thus, if the system is not invertible for $L = n$, it is never invertible.

Theorem 2.4 ([72])**.** *Consider the system* (2.6)*, where* $\mathbf{x}[k] \in \mathbb{R}^n$ *and* $\mathbf{u}[k] \in \mathbb{R}^m$*. The system is invertible with delay* L *if and only if*

$$\textit{rank}(\mathcal{J}_L) = m + \textit{rank}(\mathcal{J}_{L-1}), \tag{2.15}$$

for some $L \leq n$*, where rank*(J_{-1}) *is defined to be zero.*

It is worth noting that the upper bound on the inherent delay was improved in [94] to be $L = n - \text{nullity}(\mathbf{D}) + 1$; the proof is quite similar to the one above.

Example 2.6. Consider the F-8 aircraft given by equation (2.4), and suppose that the actuator on the aircraft could be faulty, whereby the actual input that is applied to the aircraft is different from the specified input. Mathematically, this can be modeled by setting the input to be $\mathbf{u}[k]+\mathbf{f}[k]$, where $\mathbf{u}[k]$ is the specified input, and $\mathbf{f}[k]$ is a additive error caused by the fault. The dynamics of the aircraft then become

$$\mathbf{x}[k+1] = \mathbf{A}\mathbf{x}[k] + \mathbf{B}\mathbf{u}[k] + \mathbf{B}\mathbf{f}[k].$$

where the $\mathbf{A}$ and $\mathbf{B}$ matrices are specified in (2.4). Suppose that there is a single sensor on the aircraft that measures the pitch rate, i.e.,

$$\mathbf{y}[k] = \begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix} \mathbf{x}[k].$$

Assuming that the initial state of the aircraft is known, is it possible to determine the fault input $\mathbf{f}[k]$ by looking at the output of the system? This is equivalent to asking whether the input $\mathbf{f}[k]$ is invertible. To answer this, we try to find an $L \le n$ such that (2.15) holds. For $L = 0$, we have $\mathcal{J}_0 = \mathbf{D} = \mathbf{0}$, and thus the condition is not satisfied. For $L = 1$, we have

$$\mathcal{J}_1 = \begin{bmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{CB} & \mathbf{D} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ -0.0092 & 0 \end{bmatrix},$$

which has a rank of 1. Thus, $\text{rank}(\mathcal{J}_1) - \text{rank}(\mathcal{J}_0) = 1$, and the system is invertible with delay 1.

A drawback of the above analysis is that the initial state of the system is assumed to be known, and furthermore, the state at future time-steps is obtained via the estimate of the input. However, if there is noise in the system, this may not provide an accurate representation of future states. In the next section, we will study relax the condition on knowledge of the initial state.

2.6.4 Strong Observability

While the notions of observability and invertibility deal with the separate relationships between the initial states and the output, and between the input and the output, respectively, they do not consider the relationship between the states and input (taken together) and the output. To deal with this, the following notion of *strong observability* has been established in the literature (e.g., see [51, 35, 65, 90]).

Definition 2.6 (Strong Observability)**.** A linear system of the form (2.6) is said to be *strongly observable* if, for any initial state $\mathbf{x}[0]$ and any *unknown* sequence of inputs $\mathbf{u}[0], \mathbf{u}[1], \ldots$, there is a positive integer L such that $\mathbf{x}[0]$ can be recovered from the outputs $\mathbf{y}[0], \mathbf{y}[1], \ldots, \mathbf{y}[L]$.

By the linearity of the system, the above definition is equivalent to saying that $\mathbf{y}[k] = 0$ for all k implies $\mathbf{x}[0] = 0$ (regardless of the values of the unknown inputs $\mathbf{u}[k]$).

Recall that observability and invertibility of the system could be determined by examining the observability and invertibility matrices of the system (separately); in order to characterize strong observability, we must examine the relationship between the observability and invertibility matrices. Also recall that in order for the system to be invertible, the columns of the matrix multiplying $\mathbf{u}[0]$ in (2.11) needed to be linearly independent of each other and of the columns multiplying the other unknown quantities (i.e., $\mathbf{u}[1 : L]$) in that equation. Using an identical argument, we see that the initial state $\mathbf{x}[0]$ can be recovered from (2.11) if and only if

$$\operatorname{rank}\left(\begin{bmatrix}\mathcal{O}_L & \mathcal{J}_L\end{bmatrix}\right) = n + \operatorname{rank}\left(\mathcal{J}_L\right)$$

for some nonnegative integer L; in other words, all columns of the observability matrix must be linearly independent of each other, and of all columns of the invertibility matrix.

Once again, one can ask if there is an upper bound on the number of time-steps that one would have to wait for before the above condition is satisfied (if it is satisfied at all). There is, in fact, such a bound, and the following derivation comes from [76].

First, a state $\mathbf{x}[0]$ is said to be *weakly unobservable* over $L + 1$ time-steps if there exists an input sequence $\mathbf{u}[0 : L]$ such that $\mathbf{y}[0 : L] = \mathbf{0}$. Let Σ_L denote the set of all weakly unobservable states over $L + 1$ time-steps (note that this set forms a subspace of $\mathbb{R}^n$). It is easy to see that

$$\Sigma_{L+1} \subseteq \Sigma_L \tag{2.16}$$

for all nonnegative integers L: if the state cannot be reconstructed after viewing the outputs over $L + 2$ time-steps, it cannot be reconstructed after viewing the outputs after just $L + 1$ time-steps. Next, let β denote the first nonnegative integer for which $\Sigma_\beta = \Sigma_{\beta+1}$. If $\mathbf{x}_0$ is any state in $\Sigma_{\beta+1}$, there must be an input $\mathbf{u}_0$ such that

$$\mathbf{x}_1 \triangleq \mathbf{A}\mathbf{x}_0 + \mathbf{B}\mathbf{u}_0 \in \Sigma_\beta;$$

this is because starting from $\mathbf{x}_0$, the input sequence starting with $\mathbf{u}_0$ causes the output to be zero for $\beta+2$ time-steps, and leads through the state $\mathbf{x}_1$. But, since $\Sigma_\beta = \Sigma_{\beta+1}$, we know that starting from $\mathbf{x}_1$ there is an input sequence that keeps the output zero for $\beta+2$ time-steps. This means that $\mathbf{x}_0 \in \Sigma_{\beta+2}$, because if we start from $\mathbf{x}_0$, we can apply $\mathbf{u}_0$ to go to $\mathbf{x}_1$, and then apply the input that keeps the output zero for $\beta+2$ time-steps. So we have shown that $\mathbf{x}_0 \in \Sigma_{\beta+1} \Rightarrow \mathbf{x}_0 \in \Sigma_{\beta+2}$, or equivalently $\Sigma_{\beta+1} \subseteq \Sigma_{\beta+2}$. From (2.16) we see that the opposite inclusion also holds, and so we have $\Sigma_{\beta+1} = \Sigma_{\beta+2}$. Continuing in this way, we see that

$$\Sigma_0 \supset \Sigma_1 \supset \Sigma_2 \supset \cdots \supset \Sigma_\beta = \Sigma_{\beta+1} = \cdots .$$

Since all of these spaces are subspaces of $\mathbb{R}^n$, we see that the dimension of the space can decrease at most n times, and so we have $\beta \leq n$. This leads to the following result.

Theorem 2.5. *Consider the system* (2.6) *with* $\mathbf{x}[k] \in \mathbb{R}^n$. *The system is strongly observable if and only if*

$$\textit{rank}\left(\begin{bmatrix}\mathcal{O}_L & \mathcal{J}_L\end{bmatrix}\right) = n + \textit{rank}(\mathcal{J}_L) \tag{2.17}$$

for some $L \leq n$.

Note that if the system is strongly observable, and the matrix $\begin{bmatrix}\mathbf{B} \\ \mathbf{D}\end{bmatrix}$ is full column rank, one can recover the unknown inputs as well. This is because we can obtain $\mathbf{x}[k]$ from $\mathbf{y}[k : k+L]$ for some $L \leq n$, and also $\mathbf{x}[k+1]$ from $\mathbf{y}[k+1 : k+L+1]$. Rearranging (2.6), we obtain

$$\begin{bmatrix}\mathbf{x}[k+1] - \mathbf{A}\mathbf{x}[k] \\ \mathbf{y}[k]\end{bmatrix} = \begin{bmatrix}\mathbf{B} \\ \mathbf{D}\end{bmatrix}\mathbf{u}[k],$$

and this uniquely specifies $\mathbf{u}[k]$.

Example 2.7. Consider again the F-8 from Example 2.6. To check whether one can recover the fault input $\mathbf{f}[k]$ can be recovered, regardless of the states of the system, we check whether the system is strongly observable. Specifically, for $L = n$, we have

$$\mathcal{O}_n = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0.0001 & -0.0001 & -0.8080 & 0.8942 \\ 0.0001 & -0.0004 & -1.4058 & 0.7271 \\ 0.0002 & -0.0009 & -1.7765 & 0.5240 \\ 0.0002 & -0.0015 & -1.9261 & 0.3092 \end{bmatrix},$$

$$\mathcal{J}_n = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ -0.0092 & 0 & 0 & 0 & 0 \\ 0.0028 & -0.0092 & 0 & 0 & 0 \\ 0.0125 & 0.0028 & -0.0092 & 0 & 0 \\ 0.0195 & 0.0125 & 0.0028 & -0.0092 & 0 \end{bmatrix}.$$

One can verify that $\text{rank}\left(\begin{bmatrix}\mathcal{O}_n & \mathcal{J}_n\end{bmatrix}\right) - \text{rank}(\mathcal{J}_n) = 1$, and thus the system is not strongly observable.

However, suppose that we also have a sensor that measures the velocity of the aircraft (in addition to the pitch rate). The $\mathbf{C}$ matrix would then become

$$\mathbf{C} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

and one can verify that the system is strongly observable in this case. Specifically, one can recover the fault input $\mathbf{f}[k]$ from the output of the system $\mathbf{y}[k : k+n]$ without knowing the initial state of the system.

2.6.5 System Properties and Similarity Transformations

We will now see how the properties of a given system are affected by performing a similarity transformation. Specifically, suppose we start with a particular system (2.6) and we perform a similarity transformation $\bar{\mathbf{x}} = \mathbf{T}\mathbf{x}$ to obtain a new system

$$\begin{aligned}\bar{\mathbf{x}}[k+1] &= \bar{\mathbf{A}}\bar{\mathbf{x}}[k] + \bar{\mathbf{B}}\mathbf{u}[k]\\ \mathbf{y}[k] &= \bar{\mathbf{C}}\bar{\mathbf{x}}[k] + \mathbf{D}\mathbf{u}[k]\ ,\end{aligned}$$

where $\bar{\mathbf{A}} = \mathbf{T}\mathbf{A}\mathbf{T}^{-1}$, $\bar{\mathbf{B}} = \mathbf{T}\mathbf{B}$, and $\bar{\mathbf{C}} = \mathbf{C}\mathbf{T}^{-1}$. The controllability matrix for this new realization is

$$\begin{aligned}\bar{\mathcal{C}} &= \begin{bmatrix}\bar{\mathbf{B}} & \mathbf{A}\bar{\mathbf{B}} & \cdots & \bar{\mathbf{A}}^{n-1}\bar{\mathbf{B}}\end{bmatrix}\\ &= \begin{bmatrix}\mathbf{TB} & \mathbf{TAT}^{-1}\mathbf{TB} & \cdots & (\mathbf{TAT})^{n-1}\mathbf{TB}\end{bmatrix}\\ &= \begin{bmatrix}\mathbf{TB} & \mathbf{TAB} & \cdots & \mathbf{TA}^{n-1}\mathbf{B}\end{bmatrix}\\ &= \mathbf{T}\begin{bmatrix}\mathbf{B} & \mathbf{AB} & \cdots & \mathbf{A}^{n-1}\mathbf{B}\end{bmatrix}\\ &= \mathbf{T}\mathcal{C}\ .\end{aligned}$$

Thus, the controllability matrix for the new realization is just $\mathbf{T}$ times the controllability matrix for the original realization. Recall that if $\mathbf{M}$ is a matrix and $\mathbf{T}$ is invertible, then the rank of $\mathbf{TM}$ is the same as the rank of $\mathbf{M}$ (in general, this is only true if $\mathbf{T}$ is invertible). This means that the rank of the controllability matrix for the new realization is the same as the rank of the controllability matrix for the original realization. Similarly, one can show that the rank of the observability and invertibility matrices are also unchanged. This brings us to the following result.

> Performing a similarity transformation does not change the controllability, observability, invertibility or strong observability of the system. In particular, the realization obtained from a similarity transformation is controllable/observable/invertible/strongly observable if and only if the original realization is controllable/observable/invertible/strongly observable.

Kalman Canonical Forms

Since similarity transformations do not change the properties of system, they are quite useful for analyzing system behavior. One such transformation is used to put the system into *Kalman controllability canonical form.* We will derive this transformation and form here.

First, consider the controllability matrix $\mathcal{C}_{n-1}$ for a given pair $(\mathbf{A}, \mathbf{B})$. Suppose that the system is not controllable, so that $\text{rank}(\mathcal{C}_{n-1}) = r < n$. Let $\mathbf{R}$ be an $n \times r$ matrix whose columns form a basis for the range space of $\mathcal{C}_{n-1}$ (i.e., these columns can be any set of r linearly independent columns from the controllability matrix). Define the square matrix

$$\mathbf{T} = \begin{bmatrix}\mathbf{R} & \bar{\mathbf{R}}\end{bmatrix},$$

where the $n \times (n-r)$ matrix $\bar{\mathbf{R}}$ is chosen so that $\mathbf{T}$ is invertible. Now, note that because the matrix $\mathbf{B}$ is contained in $\mathcal{C}_{n-1}$ and since all columns in the controllability matrix can be written as a linear combination of the columns in $\mathbf{R}$, we have

$$\mathbf{B} = \mathbf{R}\mathbf{B}_c = \mathbf{T}\begin{bmatrix}\mathbf{B}_c \\ \mathbf{0}\end{bmatrix}.$$

for some matrix $\mathbf{B}_c$. Similarly, recall from the derivation of the controllability index in Section 2.6.1 that $\mathbf{A}^n\mathbf{B}$ does not add any extra linearly independent columns to the controllability matrix, and so the range space of $\mathbf{A}^n\mathcal{C}_{n-1}$ is the same as the range space of $\mathcal{C}_{n-1}$. This means that $\mathbf{A}\mathbf{R} = \mathbf{R}\mathbf{A}_c$ for some matrix $\mathbf{A}_c$ (since the columns of $\mathbf{R}$ form a basis for the range space of $\mathcal{C}_{n-1}$). Using these facts, we see that

$$\begin{aligned}\bar{\mathbf{A}} \triangleq \mathbf{T}^{-1}\mathbf{A}\mathbf{T} = \mathbf{T}^{-1}\mathbf{A}\begin{bmatrix}\mathbf{R} & \bar{\mathbf{R}}\end{bmatrix} &= \mathbf{T}^{-1}\begin{bmatrix}\mathbf{A}\mathbf{R} & \mathbf{A}\bar{\mathbf{R}}\end{bmatrix} \\ &= \mathbf{T}^{-1}\begin{bmatrix}\mathbf{R}\mathbf{A}_c & \mathbf{A}\bar{\mathbf{R}}\end{bmatrix} \\ &= \begin{bmatrix}\mathbf{T}^{-1}\mathbf{R}\mathbf{A}_c & \mathbf{T}^{-1}\mathbf{A}\bar{\mathbf{R}}\end{bmatrix} \\ &= \begin{bmatrix}\mathbf{A}_c & \mathbf{A}_{12} \\ \mathbf{0} & \mathbf{A}_{\bar{c}}\end{bmatrix}, \\ \bar{\mathbf{B}} \triangleq \mathbf{T}^{-1}\mathbf{B} = \mathbf{T}^{-1}\mathbf{T}\begin{bmatrix}\mathbf{B}_c \\ \mathbf{0}\end{bmatrix} &= \begin{bmatrix}\mathbf{B}_c \\ \mathbf{0}\end{bmatrix},\end{aligned}$$

where we used the fact that $\mathbf{T}^{-1}\mathbf{R} = \mathbf{T}^{-1}\mathbf{T}\left[\begin{smallmatrix}\mathbf{I}_r \\ \mathbf{0}\end{smallmatrix}\right] = \left[\begin{smallmatrix}\mathbf{I}_r \\ \mathbf{0}\end{smallmatrix}\right]$, and defined $\mathbf{A}_{12}$ and $\mathbf{A}_{\bar{c}}$ to be the top r and bottom $n-r$ rows of $\mathbf{T}^{-1}\mathbf{A}\bar{\mathbf{R}}$, respectively. It is easy to verify that the controllability matrix for the pair $(\bar{\mathbf{A}}, \bar{\mathbf{B}})$ is given by

$$\bar{\mathcal{C}}_{n-1} = \begin{bmatrix}\mathbf{A}_c^{n-1}\mathbf{B}_c & \mathbf{A}_c^{n-2}\mathbf{B}_c & \cdots & \mathbf{A}_c\mathbf{B}_c & \mathbf{B}_c \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0}\end{bmatrix}.$$

Note that this is just the controllability matrix for the pair $(\mathbf{A}_c, \mathbf{B}_c)$ with some additional rows of zeros, and since $\text{rank}(\bar{\mathcal{C}}_{n-1}) = \text{rank}(\mathcal{C}_{n-1}) = r$, we see that this pair is controllable. Thus, the Kalman controllable canonical form for a given pair $(\mathbf{A}, \mathbf{B})$ is obtained by the pair

$$\bar{\mathbf{A}} = \begin{bmatrix}\mathbf{A}_c & \mathbf{A}_{12} \\ \mathbf{0} & \mathbf{A}_{\bar{c}}\end{bmatrix}, \quad \bar{\mathbf{B}} = \begin{bmatrix}\mathbf{B}_c \\ \mathbf{0}\end{bmatrix}, \tag{2.18}$$

where the pair $(\mathbf{A}_c, \mathbf{B}_c)$ is controllable. Note that if the original pair $(\mathbf{A}, \mathbf{B})$ is controllable to begin with, we have $\mathbf{A}_c = \mathbf{A}$ and $\mathbf{B}_c = \mathbf{B}$ in the above form.

One can also perform a similarity transformation using the *observability matrix* instead of the controllability matrix, and obtain the *Kalman observability canonical form*

$$\bar{\mathbf{A}} = \begin{bmatrix}\mathbf{A}_o & \mathbf{0} \\ \mathbf{A}_{21} & \mathbf{A}_{\bar{o}}\end{bmatrix}, \quad \bar{\mathbf{C}} = \begin{bmatrix}\mathbf{C}_o & \mathbf{0}\end{bmatrix}, \tag{2.19}$$

where the pair $(\mathbf{A}_o, \mathbf{C}_o)$ is observable. The details are similar to the derivation of the Kalman controllability canonical form, and are left as an exercise.

2.7 State-Feedback Control

We have seen how to model systems in state-space form, and to check for properties of the state-space realization. We will now see what this means for state-space control design.

It is again instructive to consider the simple scalar plant $x[k+1] = \alpha x[k] + \beta u[k]$, where $\alpha, \beta \in \mathbb{R}$, and $\beta \neq 0$. Recall from Section 2.5 that this system is stable (with $u[k] = 0$) if and only if $|\alpha| < 1$. On the other hand, if $|\alpha| > 1$, perhaps one can use the input $u[k]$ in order to prevent the system from going unstable. Specifically, suppose that we use *state-feedback* and apply $u[k] = -Kx[k]$ at each time-step, for some scalar K. The closed loop system is then $x[k+1] = (\alpha - \beta K)x[k]$, which is stable if and only if $|\alpha - \beta K| < 1$. Clearly, one can satisfy this condition with an appropriate choice of K as long as $\beta \neq 0$. In fact, choosing $K = \frac{\alpha}{\beta}$ would produce $x[k+1] = 0$, meaning that we get stability after just one time-step!

To generalize this, suppose that we have a plant with state-space model (2.6). For now, suppose that we have access to the entire state vector $\mathbf{x}[k]$ – this is not a realistic assumption in practice, because we only have access to the output vector $\mathbf{y}[k]$ (which measures a subset of the states), but let us just assume access to the full state for now. We would like use these states to construct a feedback input so that we can place the closed loop eigenvalues of the system at certain (stable) locations. We will focus on *linear state feedback* of the form

$$\mathbf{u}[k] = -\mathbf{K_1}x_1[k] - \mathbf{K_2}x_2[k] - \cdots - \mathbf{K_n}x_n[k] = -\underbrace{\begin{bmatrix}\mathbf{K}_1 & \mathbf{K}_2 & \cdots & \mathbf{K}_n\end{bmatrix}}_{\mathbf{K}}\mathbf{x}[k] \ .$$

Note that $\mathbf{u}[k]$ is a vector, in general. With this input, the closed loop state-space model becomes

$$\begin{aligned}\mathbf{x}[k+1] &= \mathbf{A}\mathbf{x}[k] + \mathbf{B}\mathbf{u}[k] = (\mathbf{A} - \mathbf{B}\mathbf{K})\,\mathbf{x}[k] \\ \mathbf{y}[k] &= (\mathbf{C} - \mathbf{D}\mathbf{K})\,\mathbf{x}[k] \ .\end{aligned}$$

The stability of this closed loop system is characterized by the eigenvalues of the matrix $\mathbf{A} - \mathbf{BK}$, and so the idea is to choose the feedback matrix $\mathbf{K}$ so that those eigenvalues are inside the unit circle (i.e., have magnitude less than 1). The following result shows when this is possible.

> It is possible to arbitrarily place the closed loop eigenvalues via state feedback of the form $\mathbf{u}[k] = -\mathbf{K}\mathbf{x}[k]$ if and only if the pair $(\mathbf{A}, \mathbf{B})$ is controllable.

The proof of necessity can be obtained by appealing to the Kalman Controllable Canonical form; the proof of sufficiency is more complicated. The above result is quite important, and we will make use of it several times, but we will omit the proof of the result here. See [95] for details.

It may be possible to find $\mathbf{K}$ such that $\mathbf{A} - \mathbf{BK}$ is stable even if the pair $(\mathbf{A}, \mathbf{B})$ is not controllable; for example, consider the case where $\mathbf{A}$ is stable, and $\mathbf{B}$ is the zero matrix.

If there is a matrix $\mathbf{K}$ such that the eigenvalues of $\mathbf{A} - \mathbf{BK}$ have magnitude less than 1, the system is said to be *stabilizable*.

Note that if a system is stabilizable but not controllable, there are some eigenvalues that cannot be placed at arbitrary locations.

Remark 2.2. If the system is controllable, the MATLAB commands `place` and `acker` can be used to find the matrix $\mathbf{K}$ such that the eigenvalues of $\mathbf{A} - \mathbf{BK}$ are at desired locations.

2.8 State Estimators and Observer Feedback

Consider again the plant (2.6). We have seen that if this realization is controllable, we can arbitrarily place the closed loop eigenvalues via state feedback of the form $\mathbf{u}[k] = -\mathbf{K}\mathbf{x}[k]$. However, there is one problem: it assumes that we have access to the entire state vector $\mathbf{x}[k]$. This is typically not the case in practice, since we only have access to the output $\mathbf{y}[k]$, which represents sensor measurements of only a few of the states. Measuring all of the states via sensors is usually not possible, since sensors can be expensive, and some states simply cannot be measured (for example, the state might represent the temperature inside an extremely hot reactor, where it is not possible to place a sensor without damaging it). How can we place the closed loop eigenvalues if we do not have access to the entire state?

The commonly used method to get around this problem is to construct an *estimator* for the state based on the output $\mathbf{y}[k]$. Specifically, the output measures some of the state variables, which are affected by the states that we do not measure. So by examining how the measured states change with time, we can potentially determine the values of the unmeasured states as well. We will do this by constructing a *state estimator* (also called a *state observer*). As one can imagine, the ability to construct such an estimator will be closely tied to the concept of *observability* that we discussed in Section 2.6.2. We will then use the state estimate $\hat{\mathbf{x}}[k]$ provided by the observer to control the system. This is called *observer feedback* and the feedback loop will look like this:

For now, we allow the observer to have access to the input $\mathbf{u}[k]$; later in the course, we will see how to build state estimators when some of the inputs to the system are unknown. Once the observer is constructed, the observer feedback input to the system is given by

$$\mathbf{u}[k] = -\mathbf{K}\hat{\mathbf{x}}[k] \ ,$$

where $\mathbf{K}$ is the same gain matrix that we would use if we had access to the actual system state (i.e., if we were using state feedback $\mathbf{u}[k] = -\mathbf{K}\mathbf{x}[k]$).

2.8.1 State Estimator Design

To see how we can obtain an estimate of the entire state, suppose that we construct a new system with state $\mathbf{z}[k]$ that *mimics* the behavior of the plant:

$$\hat{\mathbf{x}}[k+1] = \mathbf{A}\hat{\mathbf{x}}[k] + \mathbf{B}\mathbf{u}[k] \ .$$

If we initialize this system with $\hat{\mathbf{x}}[0] = \mathbf{x}[0]$ and we apply the same input $\mathbf{u}[k]$ to this system and the plant, we would have $\hat{\mathbf{x}}[k] = \mathbf{x}[k]$ for all time. Thus, we would have a perfect estimate of the state for all time, and we could use the state feedback control $\mathbf{u}[k] = -\mathbf{K}\hat{\mathbf{x}}[k]$, where $\mathbf{K}$ is the control gain required to place the eigenvalues at desired locations. In summary, if we knew the initial state $\mathbf{x}[0]$ of the system, we could technically obtain an estimate of the state at any time. However, there are some problems with this:

- We may not know the initial state of the system (especially if we cannot measure some of the states of the system).
- The above observer does not make use of any measurements of the states, and thus it has no way of correcting itself if the estimated states start diverging from the actual states (e.g., due to noise or disturbances in the system).

In order to fix these shortcomings, we will modify the observer equation as follows:

$$\hat{\mathbf{x}}[k+1] = \mathbf{A}\hat{\mathbf{x}}[k] + \mathbf{B}\mathbf{u}[k] + \mathbf{L}(\mathbf{y}[k] - \mathbf{C}\hat{\mathbf{x}}[k] - \mathbf{D}\mathbf{u}[k]) \ . \tag{2.20}$$

In this modified observer, the role of the *corrective term* $\mathbf{L}(\mathbf{y}[k] - \mathbf{C}\hat{\mathbf{x}}[k] - \mathbf{D}\mathbf{u}[k])$ is to utilize the measurements of the state vector in order to help the observer do a good job of tracking the state. Specifically, since $\mathbf{y}[k] = \mathbf{C}\mathbf{x}[k] + \mathbf{D}\mathbf{u}[k]$, the term $\mathbf{y}[k] - \mathbf{C}\hat{\mathbf{x}}[k] - \mathbf{D}\mathbf{u}[k]$ represents the error between the measured states and the estimates of those states. If $\hat{\mathbf{x}}[k] = \mathbf{x}[k]$ (i.e., the state observer is perfectly synchronized with the state), then the term $\mathbf{y}[k] - \mathbf{C}\hat{\mathbf{x}}[k] - \mathbf{D}\mathbf{u}[k]$ will be zero. If the state estimate is different from the actual state, however, the hope is that the term $\mathbf{y}[k] - \mathbf{C}\hat{\mathbf{x}}[k] - \mathbf{D}\mathbf{u}[k]$ will also be nonzero, and help to reduce the estimation error to zero. The gain matrix $\mathbf{L}$ is used to ensure that this will happen.

To see how to choose $\mathbf{L}$, let us examine the *estimation error* defined as $\mathbf{e}[k] = \mathbf{x}[k] - \hat{\mathbf{x}}[k]$. The evolution of the estimation error is given by

$$\begin{aligned}
\mathbf{e}[k+1] &= \mathbf{x}[k+1] - \hat{\mathbf{x}}[k+1] \\
&= \mathbf{A}\mathbf{x}[k] + \mathbf{B}\mathbf{u}[k] - \mathbf{A}\hat{\mathbf{x}}[k] - \mathbf{B}\mathbf{u}[k] - \mathbf{L}(\mathbf{y}[k] - \mathbf{C}\hat{\mathbf{x}}[k] - \mathbf{D}\mathbf{u}[k]) \\
&= \mathbf{A}(\mathbf{x}[k] - \hat{\mathbf{x}}[k]) - \mathbf{L}(\mathbf{C}\mathbf{x}[k] - \mathbf{C}\hat{\mathbf{x}}[k]) \\
&= \mathbf{A}\mathbf{e}[k] - \mathbf{L}\mathbf{C}\mathbf{e}[k] \\
&= (\mathbf{A} - \mathbf{L}\mathbf{C})\mathbf{e}[k] \ .
\end{aligned}$$

This is simply an autonomous linear system, and if we would like the estimation error to go to zero regardless of the initial estimation error, we have to choose the matrix $\mathbf{L}$ so that the eigenvalues of $\mathbf{A} - \mathbf{L}\mathbf{C}$ all have magnitude less than 1.

Condition For Placing Eigenvalues of $\mathbf{A} - \mathbf{L}\mathbf{C}$: Observability

To determine conditions on $\mathbf{A}$ and $\mathbf{C}$ which will allow us to arbitrarily place the eigenvalues of $\mathbf{A} - \mathbf{L}\mathbf{C}$, let us make a connection to controllability. Recall that if the pair ($\mathbf{A}$, $\mathbf{B}$) is controllable, then it is possible to place the eigenvalues of $\mathbf{A} - \mathbf{B}\mathbf{K}$ arbitrarily via a choice of matrix $\mathbf{K}$. For the observer, we are dealing with the matrix $\mathbf{A} - \mathbf{L}\mathbf{C}$; this is different from $\mathbf{A} - \mathbf{B}\mathbf{K}$ because the gain matrix $\mathbf{L}$ pre-multiplies the matrix $\mathbf{C}$, whereas the gain matrix $\mathbf{K}$ post-multiplies the matrix $\mathbf{B}$. However, note that the eigenvalues of a matrix are the same as the eigenvalues of the transpose of the matrix. This means that the eigenvalues of $\mathbf{A} - \mathbf{L}\mathbf{C}$ are the same as the eigenvalues of the matrix $\mathbf{A}' - \mathbf{C}'\mathbf{L}'$, and this matrix has the same form as $\mathbf{A} - \mathbf{B}\mathbf{K}$. Based on our discussion of controllability, we know that if the pair $(\mathbf{A}', \mathbf{C}')$ is controllable, then we can choose $\mathbf{L}$ to place the eigenvalues of $\mathbf{A}' - \mathbf{C}'\mathbf{L}'$ (and thus $\mathbf{A} - \mathbf{L}\mathbf{C}$) at arbitrary locations. Recall that the pair $(\mathbf{A}', \mathbf{C}')$ is controllable if and only if the pair $(\mathbf{A}, \mathbf{C})$ is observable, which brings us to the following result.

> The eigenvalues of $\mathbf{A} - \mathbf{L}\mathbf{C}$ can be placed arbitrarily if and only if the pair $(\mathbf{A}, \mathbf{C})$ is observable. If there exists a matrix $\mathbf{L}$ such that $\mathbf{A} - \mathbf{L}\mathbf{C}$ is stable, the pair $(\mathbf{A}, \mathbf{C})$ is said to be *detectable*.

2.8.2 The Separation Principle

In the last section, we designed the observer (2.20) and used the observer feedback input $\mathbf{u}[k] = -\mathbf{K}\hat{\mathbf{x}}[k]$ where $\mathbf{K}$ was chosen to place the eigenvalues of $\mathbf{A} - \mathbf{B}\mathbf{K}$ at desired locations. Note that we chose the gain $\mathbf{K}$ *independently* of the observer – we pretended that we were just using full state feedback, instead of an estimate of the state. To make sure that this is valid, let us examine the equations for the entire closed loop system:

$$\begin{aligned}\mathbf{x}[k+1] &= \mathbf{A}\mathbf{x}[k] + \mathbf{B}\mathbf{u}[k]\\ &= \mathbf{A}\mathbf{x}[k] - \mathbf{B}\mathbf{K}\hat{\mathbf{x}}[k]\\ \hat{\mathbf{x}}[k+1] &= \mathbf{A}\hat{\mathbf{x}}[k] + \mathbf{B}\mathbf{u}[k] + \mathbf{L}(\mathbf{y}[k] - \mathbf{C}\hat{\mathbf{x}} - \mathbf{D}\mathbf{u}[k])\\ &= (\mathbf{A} - \mathbf{B}\mathbf{K} - \mathbf{L}\mathbf{C})\hat{\mathbf{x}}[k] + \mathbf{L}\mathbf{C}\mathbf{x}[k]\ .\end{aligned}$$

The closed loop system therefore has $2n$ states (n states corresponding to the plant, and n states corresponding to the observer). In matrix-vector form, this is

$$\begin{bmatrix}\mathbf{x}[k+1]\\ \hat{\mathbf{x}}[k+1]\end{bmatrix} = \underbrace{\begin{bmatrix}\mathbf{A} & -\mathbf{B}\mathbf{K}\\ \mathbf{L}\mathbf{C} & \mathbf{A}-\mathbf{B}\mathbf{K}-\mathbf{L}\mathbf{C}\end{bmatrix}}_{\mathbf{A}_{cl}}\begin{bmatrix}\mathbf{x}[k]\\ \hat{\mathbf{x}}[k]\end{bmatrix}\ .$$

The closed loop poles are given by the eigenvalues of the matrix $\mathbf{A}_{cl}$. It is hard to determine the eigenvalues of $\mathbf{A}_{cl}$ from its current form, but recall that the eigenvalues of a matrix are unchanged if we perform a similarity transformation on it. Let us define the similarity transformation matrix

$$\mathbf{T} = \begin{bmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{I} & -\mathbf{I}\end{bmatrix}\ ,$$

which has the property that $\mathbf{T} = \mathbf{T}^{-1}$. We then have

$$\begin{aligned}\text{eig}(\mathbf{A}_{cl}) &= \text{eig}\left(\mathbf{T}\mathbf{A}_{cl}\mathbf{T}^{-1}\right)\\ &= \text{eig}\left(\begin{bmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{I} & -\mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{A} & -\mathbf{B}\mathbf{K}\\ \mathbf{L}\mathbf{C} & \mathbf{A}-\mathbf{B}\mathbf{K}-\mathbf{L}\mathbf{C}\end{bmatrix}\begin{bmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{I} & -\mathbf{I}\end{bmatrix}\right)\\ &= \text{eig}\left(\begin{bmatrix}\mathbf{A}-\mathbf{B}\mathbf{K} & \mathbf{B}\mathbf{K}\\ \mathbf{0} & \mathbf{A}-\mathbf{L}\mathbf{C}\end{bmatrix}\right)\ .\end{aligned}$$

Now, we use the fact that the eigenvalues of a matrix of the form $\begin{bmatrix}\mathbf{M}_1 & \mathbf{M}_2\\ 0 & \mathbf{M}_3\end{bmatrix}$ (where $\mathbf{M}_1$ and $\mathbf{M}_3$ are square matrices) are just the eigenvalues of matrix $\mathbf{M}_1$ together with the eigenvalues of matrix $\mathbf{M}_3$. This means that the eigenvalues of $\mathbf{A}_{cl}$ are simply the eigenvalues of $\mathbf{A} - \mathbf{B}\mathbf{K}$ together with the eigenvalues of $\mathbf{A} - \mathbf{L}\mathbf{C}$. In other words, the closed loop eigenvalues are the state feedback eigenvalues along with the eigenvalues of the observer – this shows that we can, in fact, design the feedback gain $\mathbf{K}$ independently of the observer.

2.9 Matrix Pencil Characterizations of System Properties

While the properties of controllability, observability, invertibility and strong observability were characterized by directly examining the controllability, observability and invertibility matrices, there are alternative characterizations that are also possible. We will discuss one such characterization here that has the appealing feature of not depending on powers of the matrix $\mathbf{A}$ (which appears in all of the matrices described above).

2.9.1 Controllability and Observability

Theorem 2.6. *Consider the system* (2.6) *with* $\mathbf{x}[k] \in \mathbb{R}^n$, $\mathbf{u}[k] \in \mathbb{R}^m$ *and* $\mathbf{y}[k] \in \mathbb{R}^p$.

1. *The pair* $(\mathbf{A}, \mathbf{B})$ *is controllable (stabilizable) if and only if rank* $\left(\begin{bmatrix}\mathbf{A} - z\mathbf{I}_n & \mathbf{B}\end{bmatrix}\right) = n$ *for all* $z \in \mathbb{C}$ *(*$|z| \geq 1$*).*
2. *The pair* $(\mathbf{A}, \mathbf{C})$ *is observable (detectable) if and only if rank* $\left(\begin{bmatrix}\mathbf{A} - z\mathbf{I}_n \\ \mathbf{C}\end{bmatrix}\right) = n$ *for all* $z \in \mathbb{C}$*, (*$|z| \geq 1$*).*

Proof. We will prove the controllability result; the observability result can be obtained in a similar manner.

(Sufficiency:) First, we will prove that

$$\text{rank}\left(\begin{bmatrix}\mathbf{A} - z\mathbf{I}_n & \mathbf{B}\end{bmatrix}\right) = n \ \forall \ z \in \mathbb{C} \Rightarrow \text{rank}(\mathcal{C}_{n-1}) = n.$$

Recall from Section 2.6.5 that if the rank of the controllability matrix is $r < n$ (i.e., it is not controllable), there exists a matrix $\mathbf{T}$ (constructed from r linearly independent columns from the controllability matrix) such that

$$\mathbf{T}^{-1}\mathbf{A}\mathbf{T} = \begin{bmatrix}\mathbf{A}_c & \mathbf{A}_{12} \\ \mathbf{0} & \mathbf{A}_{\bar{c}}\end{bmatrix}, \ \mathbf{T}^{-1}\mathbf{B} = \begin{bmatrix}\mathbf{B}_c \\ \mathbf{0}\end{bmatrix},$$

where $\mathbf{A}_c$ is an $r \times r$ matrix and $\mathbf{A}_{\bar{c}}$ is an $(n-r) \times (n-r)$ matrix. Next, note that multiplying a matrix on the left and right by invertible matrices does not change the rank of the matrix. Thus,

$$\begin{aligned}\text{rank}\left(\begin{bmatrix}\mathbf{A} - z\mathbf{I}_n & \mathbf{B}\end{bmatrix}\right) &= \text{rank}\left(\mathbf{T}^{-1}\begin{bmatrix}\mathbf{A} - z\mathbf{I}_n & \mathbf{B}\end{bmatrix}\begin{bmatrix}\mathbf{T} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{n-r}\end{bmatrix}\right) \\ &= \text{rank}\left(\begin{bmatrix}\mathbf{T}^{-1}\mathbf{A}\mathbf{T} - z\mathbf{I}_n & \mathbf{T}^{-1}\mathbf{B}\end{bmatrix}\right) \\ &= \text{rank}\left(\begin{bmatrix}\mathbf{A}_c - z\mathbf{I}_r & \mathbf{A}_{12} & \mathbf{B}_c \\ \mathbf{0} & \mathbf{A}_{\bar{c}} - z\mathbf{I}_{n-r} & \mathbf{0}\end{bmatrix}\right).\end{aligned}$$

This matrix will have rank n if and only if all of its rows are linearly independent (because there are only n rows in the matrix). In particular, this means that all of bottom $n-r$ rows of the matrix must be linearly independent, which is equivalent to saying that all of the rows of the matrix $\mathbf{A}_{\bar{c}} - z\mathbf{I}_{n-r}$ must be linearly independent. However, whenever $n - r > 0$, we can choose z to be an eigenvalue of $\mathbf{A}_{\bar{c}}$, and this would cause these rows to be linearly dependent (since $\det(\mathbf{A}_{\bar{c}} - z\mathbf{I}_{n-r}) = 0$ for any eigenvalue z). Thus, the only way for $\begin{bmatrix}\mathbf{A} - z\mathbf{I}_n & \mathbf{B}\end{bmatrix}$ to have rank n for all z is if $r = n$ (i.e., the system is controllable).

(Necessity:) Next, we will prove that

$$\text{rank}(\mathcal{C}_{n-1}) = n \Rightarrow \text{rank}\left(\begin{bmatrix}\mathbf{A} - z\mathbf{I}_n & \mathbf{B}\end{bmatrix}\right) = n \ \forall \ z \in \mathbb{C}.$$

We will do this by proving the *contrapositive*, that is,

$$\exists z_0 \in \mathbb{C} \text{ s.t. } \operatorname{rank}\left(\begin{bmatrix}\mathbf{A} - z_0\mathbf{I}_n & \mathbf{B}\end{bmatrix}\right) < n \Rightarrow \operatorname{rank}(\mathcal{C}_{n-1}) < n.$$

To do this, note that if $\operatorname{rank}\left(\begin{bmatrix}\mathbf{A} - z_0\mathbf{I}_n & \mathbf{B}\end{bmatrix}\right) < n$, then the rows of this matrix form a linearly dependent set. Thus, there exists a row vector $\mathbf{v}'$ such that

$$\mathbf{v}'\begin{bmatrix}\mathbf{A} - z_0\mathbf{I}_n & \mathbf{B}\end{bmatrix} = \mathbf{0},$$

or equivalently

$$\mathbf{v}'\mathbf{A} = z_0\mathbf{v}', \;\; \mathbf{v}'\mathbf{B} = \mathbf{0}.$$

This means that $\mathbf{v}'$ must be a left eigenvector of $\mathbf{A}$, corresponding to eigenvalue z_0. This means that $\mathbf{v}'\mathbf{A}^k = z_0^k\mathbf{v}'$ for any $k \in \mathbb{N}$, and thus

$$\begin{aligned}\mathbf{v}'&\begin{bmatrix}\mathbf{A}^{n-1}\mathbf{B} & \mathbf{A}^{n-2}\mathbf{B} & \cdots & \mathbf{A}\mathbf{B} & \mathbf{B}\end{bmatrix}\\ &= \begin{bmatrix}z_0^{n-1}\mathbf{v}'\mathbf{B} & z_0^{n-2}\mathbf{v}'\mathbf{B} & \cdots & z_0\mathbf{v}'\mathbf{B} & \mathbf{v}'\mathbf{B}\end{bmatrix} = \mathbf{0}.\end{aligned}$$

Thus the rows of $\mathcal{C}_{n-1}$ are also linearly dependent, which means that

$$\operatorname{rank}(\mathcal{C}_{n-1}) < n.$$

□

The tests for controllability and observability in Theorem 2.6 are called the *Popov-Belevitch-Hautus* tests. Further discussion of these tests can be found in [1, 36].

2.9.2 Invertibility

To characterize alternative conditions for invertibility and strong observability, it will be useful to introduce the following terminology.

Definition 2.7 (Matrix Pencil)**.** For the linear system (2.6), the matrix

$$\mathbf{P}(z) = \begin{bmatrix}\mathbf{A} - z\mathbf{I}_n & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{bmatrix}$$

is called the *matrix pencil* of the set $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$. The variable z is an element of $\mathbb{C}$.

Theorem 2.7. *Consider the system* (2.6) *with* $\mathbf{x}[k] \in \mathbb{R}^n$, $\mathbf{u}[k] \in \mathbb{R}^m$ *and* $\mathbf{y}[k] \in \mathbb{R}^p$. *The system is invertible if and only if*

$$\text{rank}\left(\begin{bmatrix} \mathbf{A} - z\mathbf{I}_n & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}\right) = n + m$$

for at least one $z \in \mathbb{C}$.

Proof. To prove this result, we will make use of the transfer function of the system. Specifically, recall from Section 2.4.1 that the z-transform of system (2.6) yields the following input-output representation:

$$\mathbf{Y}(z) = \mathbf{C}(z\mathbf{I}_n - \mathbf{A})^{-1}\mathbf{x}[0] + \underbrace{(\mathbf{C}(z\mathbf{I}_n - \mathbf{A})^{-1})\mathbf{B} + \mathbf{D})}_{\mathbf{H}(z)}\mathbf{U}(z).$$

The transfer function matrix $\mathbf{H}(z)$ captures how the input to the system is reflected in the output, and thus the system is invertible if and only if the transfer function has rank m (over the field of all rational functions of z). We will now relate the rank of the transfer function matrix to the rank of $\mathbf{P}(z)$, and to do this, we use the following trick from [91]:

$$\begin{bmatrix} \mathbf{A} - z\mathbf{I}_n & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_n & \mathbf{0} \\ \mathbf{C}(\mathbf{A} - z\mathbf{I}_n)^{-1} & \mathbf{I}_p \end{bmatrix} \begin{bmatrix} \mathbf{A} - z\mathbf{I}_n & \mathbf{0} \\ \mathbf{0} & \mathbf{C}(z\mathbf{I}_n - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{I}_n & (\mathbf{A} - z\mathbf{I}_n)^{-1}\mathbf{B} \\ \mathbf{0} & \mathbf{I}_m \end{bmatrix}.$$

Since the left-most and right-most matrices on the right-hand-side of the above equation are invertible, we have

$$\begin{aligned} \text{rank}\left(\begin{bmatrix} \mathbf{A} - z\mathbf{I}_n & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}\right) &= \text{rank}\left(\begin{bmatrix} \mathbf{A} - z\mathbf{I}_n & \mathbf{0} \\ \mathbf{0} & \mathbf{H}(z) \end{bmatrix}\right) \\ &= \text{rank}\left(\mathbf{A} - z\mathbf{I}_n\right) + \text{rank}\left(\mathbf{H}(z)\right). \end{aligned}$$

If the system is invertible, then $\text{rank}\left(\mathbf{H}(z)\right) = m$ over the field of rational functions of z, and this means that it will have rank m for almost any numerical choice of $z \in \mathbb{C}$. Furthermore, $\text{rank}\left(\mathbf{A} - z\mathbf{I}_n\right) = n$ for any z that is not an eigenvalue of $\mathbf{A}$. Thus, if the system is invertible, we have $\text{rank}\left(\mathbf{P}(z)\right) = n + m$ for almost any choice of z. On the other hand, if the system is not invertible, then $\text{rank}(\mathbf{H}(z)) < m$ for every value of z, and thus $\text{rank}(\mathbf{P}(z)) < n + m$ for every value of z. □

2.9.3 Strong Observability

To characterize strong observability of a system, we will assume without loss of generality that $\text{rank}\left[\begin{smallmatrix}\mathbf{B}\\\mathbf{D}\end{smallmatrix}\right] = m$ (otherwise, we can just use those columns of the matrix that are linearly independent, and still be able to affect the system with the inputs in the same way).

Definition 2.8 (Invariant Zero)**.** The complex number $z_0 \in \mathbb{C}$ is called an *invariant zero* of the system (2.6) if $\text{rank}\,(\mathbf{P}(z_0)) < n+m$.

Theorem 2.8. *Consider the system* (2.6) *with* $\mathbf{x}[k] \in \mathbb{R}^n$*,* $\mathbf{u}[k] \in \mathbb{R}^m$ *and* $\mathbf{y}[k] \in \mathbb{R}^p$*. The system is strongly observable if and only if*

$$\mathit{rank}\left(\begin{bmatrix}\mathbf{A} - z\mathbf{I}_n & \mathbf{B} \\ \mathbf{C} & \mathbf{D}\end{bmatrix}\right) = n + m$$

for all $z \in \mathbb{C}$ *(i.e., the system has no invariant zeros).*

Proof. (Necessity:) Suppose that $\text{rank}(\mathbf{P}(z_0)) < n + m$ for some $z_0 \in \mathbb{C}$. Then, there exists a vector $\mathbf{v}$ in the null-space of $\mathbf{P}(z_0)$. Denote the top n components of $\mathbf{v}$ by $\mathbf{x}_0$ and the bottom m components by $\mathbf{u}_0$. This means that

$$\begin{bmatrix}\mathbf{A} - z_0\mathbf{I}_n & \mathbf{B} \\ \mathbf{C} & \mathbf{D}\end{bmatrix}\begin{bmatrix}\mathbf{x}_0 \\ \mathbf{u}_0\end{bmatrix} = \mathbf{0} \Leftrightarrow \begin{matrix}\mathbf{A}\mathbf{x}_0 + \mathbf{B}\mathbf{u}_0 = z_0\mathbf{x}_0 \\ \mathbf{C}\mathbf{x}_0 + \mathbf{D}\mathbf{u}_0 = \mathbf{0}\end{matrix} \tag{2.21}$$

This indicates that if the initial state of the system is $\mathbf{x}_0$ and we apply the input $\mathbf{u}[0] = \mathbf{u}_0$, then $\mathbf{x}[1] = z_0\mathbf{x}_0$ and $\mathbf{y}[0] = \mathbf{0}$. Next, multiply (2.21) by z_0 on the right to obtain

$$\begin{bmatrix}\mathbf{A} - z_0\mathbf{I}_n & \mathbf{B} \\ \mathbf{C} & \mathbf{D}\end{bmatrix}\begin{bmatrix}z_0\mathbf{x}_0 \\ z_0\mathbf{u}_0\end{bmatrix} = \mathbf{0} \Leftrightarrow \begin{matrix}\mathbf{A}(z_0\mathbf{x}_0) + \mathbf{B}(z_0\mathbf{u}_0) = z_0^2\mathbf{x}_0 \\ \mathbf{C}(z_0\mathbf{x}_0) + \mathbf{D}(z_0\mathbf{u}_0) = \mathbf{0}\end{matrix}$$

This means that if we applied the input $\mathbf{u}[1] = z_0\mathbf{u}_0$, the state at time-step 2 would be $\mathbf{x}[2] = z_0^2\mathbf{x}_0$ and $\mathbf{y}[1] = 0$. Continuing in this way, we see that if we apply the input sequence $\mathbf{u}[k] = z_0^k\mathbf{u}_0$ and the initial state of the system is $\mathbf{x}_0$, the state of the system for all time-steps will satisfy $\mathbf{x}[k] = z_0^k\mathbf{x}_0$ and the output of the system will be $\mathbf{y}[k] = \mathbf{0}$. This is the same output that is obtained when the initial state of the system is $\mathbf{x}[0] = \mathbf{0}$ and $\mathbf{u}[k] = \mathbf{0}$ for all k, and so we cannot recover $\mathbf{x}_0$ from the output of the system.

(Sufficiency:) The proof of sufficiency is given in [76], and we will not cover it here. □

The proof of necessity shows that if the system has an invariant zero z_0, there is an initial state and a set of inputs such that the output is zero for all time, but the state gets multiplied by z_0 at each time-step. Now if $|z_0| < 1$, this may not be of major concern, since the state will simply decay to zero. On the other hand, if $|z_0| \geq 1$, the state will explode (or stay constant), and we would never be able to determine this from the output. Just as we discussed the notions of detectability and stabilizability as relaxations of observability and controllability, respectively, we can also consider the notion of *strong detectability* to capture the less serious of these cases.

Definition 2.9. The linear system (2.6) is *strongly detectable* if $\mathbf{y}[k] = \mathbf{0}$ for all k implies that $\mathbf{x}[k] \to \mathbf{0}$.

In words, the above definition says that any initial state that holds the output equal to zero for all time (and thus cannot be differentiated from the all zero initial state) must eventually decay to zero. The matrix pencil test in Theorem 2.8 can be generalized to accommodate this case as follows.

Theorem 2.9. *The system* (2.6) *is strongly detectable if and only if all invariant zeros have magnitude less than* 1*.*

From the above theorems, note that a system can be strongly observable, strongly detectable, or invertible only if the number of outputs p is at least as large as the number of inputs m. Otherwise, the matrix pencil $\mathbf{P}(z)$ will have more columns than rows, and the maximum rank could be no greater than $n + p$, which is less than $n + m$.