An Introduction to System Identification

Matthew Hölzel
hoelzel@uni-bremen.de
University of Bremen
28359 Bremen, Germany

Contents

PART I

Introduction

SECTION 1

Introduction

System identification is the process of determining a model to describe the input/output relationship of a system given some observed inputs and outputs. Usually the first step will be to choose or guess a model structure, after which the coefficients of the model will be chosen to minimize some cost function. However, the structure of the model might also be known *a priori*, such as when a model is derived using first principles (e.g. Newton's laws). In such a case, the *system identification* problem is sometimes called a *parameter identification* problem.

Example 1.1 Suppose we want to model the position y of a mass m sliding on a smooth, frictionless surface (such as ice) subject to the external force u, as shown in Figure 1. For instance, we might need such a model if we want to control the position y using the input u, where we regard y as the *output* of the system.

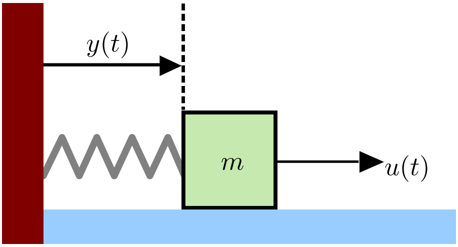


Figure 1: A mass m sliding on a smooth frictionless surface subject to the external force u, where the mass is attached to a rigid wall via a spring, and y denotes the distance between the mass and the wall.

If we have no idea how to model the mass-spring system in Figure 1, then we need to guess a generic model structure that we think is capable of describing the dynamic relationship between the input u and output y. For instance, since ordinary differential equations are so prevalent in physics problems, we might guess that our system can be approximately modelled by a second-order ordinary differential equation, that is,

$$\ddot{y}(t) + \alpha_1 \dot{y}(t) + \alpha_0 y(t) \approx \beta_2 \ddot{u}(t) + \beta_1 \dot{u}(t) + \beta_0 u(t) \tag{1.1}$$

although at this point, we have no idea whether (1.1) is a good representation of the system, and we have not chosen the specific parameters α_i and β_i in the model (1.1).

Next, we are going to try to choose the parameters α_i and β_i so that the model (1.1) best describes our system, and to do that, we need data. Specifically, we need to measure the output y in response to some known input u, which we can then plug into the model (1.1). For instance, if we measure the inputs and outputs (along with their first and second derivatives) at times $t_1, \dots, t_N$, then the *error* e in our model at time t_i, when using the parameters $\alpha_0, \alpha_1, \beta_0, \beta_1$, and β_2, is given by

$$\begin{aligned} e\big(t_i,&\alpha_0, \alpha_1, \beta_0, \beta_1, \beta_2\big) \\ &= \ddot{y}(t_i) + \alpha_1 \dot{y}(t_i) + \alpha_0 y(t_i) - \Big[\beta_2 \ddot{u}(t_i) + \beta_1 \dot{u}(t_i) + \beta_0 u(t_i)\Big] \end{aligned} \tag{1.2}$$

Remark 1.1 There are "smart" ways of choosing good input signals, but for now, you could just imagine that we hit the mass with a hammer with some known force at some known time, and just measure how the mass moves in response to our violent aggression. □

Since smaller errors imply that the model (1.1) better fits the data, we are therefore going to choose the parameters α_i and β_i which minimize the errors $e(t_1), \dots, e(t_N)$ in some sense. For instance, we could choose α_i and β_i to minimize the least-squares cost function

$$\sum_{i=1}^{N} e^2\big(t_i, \alpha_0, \alpha_1, \beta_0, \beta_1, \beta_2\big) \tag{1.3}$$

the least-absolute value cost function

$$\sum_{i=1}^{N} \Big| e\big(t_i, \alpha_0, \alpha_1, \beta_0, \beta_1, \beta_2\big)\Big| \tag{1.4}$$

or some other function of the errors, although it is important to note that the parameters α_i and β_i which minimize (1.3) will generally be different than those that minimize (1.4), that is, the values that we choose to "best describe" the system will depend heavily on the cost function which we choose to minimize.

Remark 1.2 There is generally no *right* or *wrong* cost function. Rather, the cost function which we will try to minimize is usually chosen because it is easy to minimize and produces reasonable parameter estimates. □

If, on the other hand, we knew how to exactly model the system from first principles, and we could do it easily, then we should almost always do that instead. For instance, if we knew about Hooke's law, then instead of (1.1), we could use the model

$$m\ddot{y}(t) = u(t) - k\big[y(t) - y_0\big] \tag{1.5}$$

where k denotes the spring constant, and y_0 denotes the distance y at which the spring is unstretched. In this case:

1) We would know that our model is capable of describing the system's dynamics.

2) We would only have to deal with the reduced set of unknown parameters m, k, and y_0.

Specifically, we would still have to minimize a cost function like (1.3) or (1.4), but we would only have to perform the optimization over m, k, and y_0, since the errors would now be of the form

$$e\big(t_i, m, k, y_0\big) = m\ddot{y}(t_i) - \Big(u(t_i) - k\big[y(t_i) - y_0\big]\Big) \tag{1.6}$$

□

Example 1.1 demonstrates a typical system identification application in which we choose a model, gather data, and then choose the model parameters to fit the data. However, this is just the tip of the iceberg. System identification also attempts to answer questions such as:

(i) *How accurate is the estimated model?*

(ii) *How do sensor and process noise affect the estimated model?*

(iii) *How does the criterion used for estimating the coefficients, such as* (1.3)*, affect their accuracy?*

(iv) *If we use more measurements, or a completely different data set, how would the estimated model change?*

(v) *If we assume or guess a model that is structurally different from the real system, then can we still expect o model to be somehow meaningful?*

Some common systems to which we could apply system identification are shown in Figure 2, along with some inputs and outputs that we might want to model. For instance, we might want to model the dynamics between an airplane's thrust and lift, or model the dynamics between a car's gas pedal and its acceleration.

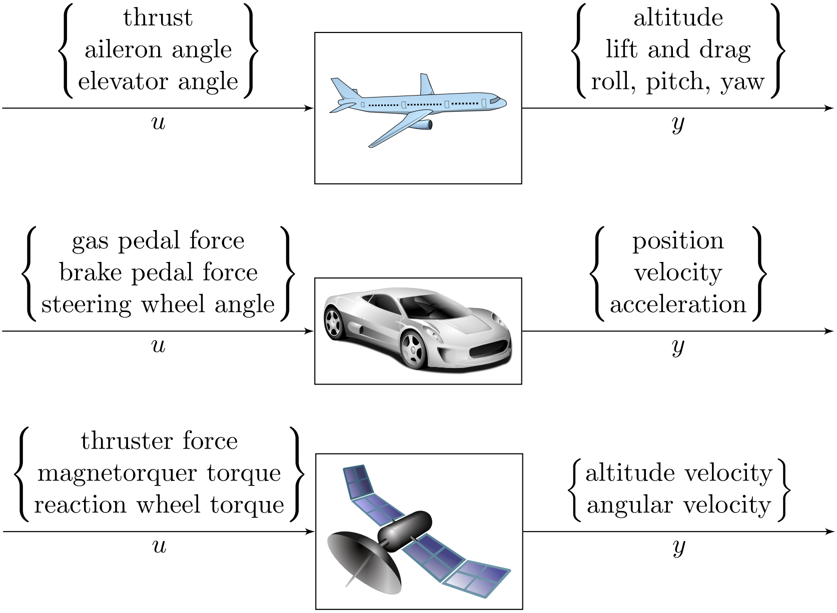


Figure 2: Some common systems to which system identification is applied.

Furthermore, some common cases where system identification is used are:

- *When a model is hard or impossible to derive from first principles, either due to the complexity of the system, or because there does not exist a traditional method of modeling the system.*

  Example:

  Suppose we need to model the reaction of a satellite with several solar panels to a thruster being fired. For instance, we might need such a model for the purpose of controlling the satellite. Unfortunately, modeling flexible systems like solar panels is a serious challenge, where damping is very difficult to predict. If we start modeling the cables inside the satellite, then this task only gets more difficult. On the other hand, it might be relatively easy for us to test the response of the satellite in the laboratory to various shakes and twists, after which we can performing system identification to generate a model.

- *When the system itself is not known precisely enough to be modeled accurately from first principles.*

  Example: Suppose we design a medical robot to perform surgery. Then clearly the movements of the robot's arms need to be very precise. However, even if we develop an extremely accurate analytic model for the robot's movements, there will always be some difference between the model and the real robot due to the manufacturing process. In this case, the additional manufacturing errors must be corrected away in another manner such as with system identification (*calibration* is a special case). For instance, even the problem of determining the robot's weight could be construed as a system identification problem.

- *When a system changes, and a new model is quickly needed to regain control of the system.*

  Example:

  Suppose we have a self-driving car, that is, a car with something like an airplane's autopilot. The autopilot can be robustly designed in the factory to be able to deal with bumps in the roads, turns, etc. However, when the system changes abruptly while driving, the autopilot needs to be able to adjust its algorithm in real-time. For instance, how should the autopilot respond to a tire falling off, or to tires gradually wearing out over time? Clearly, the autopilot needs to respond differently to these two events. The first step toward regaining control of the system is to apply system identification to identify a new system model, after which the autopilot can determine the appropriate actions to take.

- *When we need a simplified model of the system for some other purposes, like control.*

  Example:

  Suppose we have an extremely detailed finite-element model of an airplane. Many engineers spent many years developing very accurate models to make sure that the airplane will produce enough lift and the engines will provide enough thrust. Perhaps surprisingly, however, the autopilot only needs to know some very coarse-grained information about the airplane's dynamics. Specifically, it might only need to know a few coefficients such as the drag coefficient, which represents the relationship between the amount of drag on the airplane, and the angle that the flaps are set to. Such a simple coefficient could easily be determined using system identification techniques while a real pilot is flying the plane.

Although system identification can be applied to a wide range of applications in a wide range of fields, it should never be applied blindly. The practitioner should always know something about the underlying system he is trying to identify and the data that he is working with. For instance:

- Do you expect the dynamics to be linear, or do you expect to see some other phenomenon like hysteresis?
- Were the inputs measured or are they known exactly?
- Were the outputs measured with additive or multiplicative noise?
- Can you characterize the type of noise (white or colored, Gaussian or logarithmic)?

Hence model selection and data analysis are central themes in system identification.

PART II

Systems and Models

A *system*, in the context of this book, is loosely defined as an object which accepts inputs u, produces outputs y, and has the notion of a *state* (which we will introduce later in Definition 4.1). Hence a *system* can refer to either a mathematical object, such as a differential equation, or a real object, such as a car. A *model*, on the other hand, is a mathematical object which is used to describe a system.

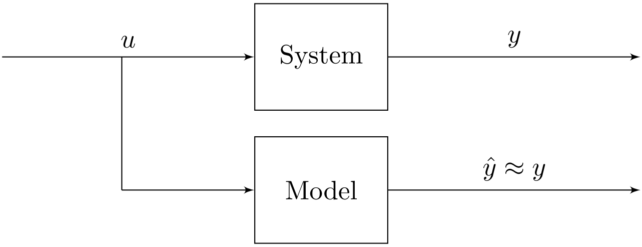


Figure 3: The input u and output y of a system, along with a model of the system which produces the modeled output $\hat{y}$.

When a system refers to a real object, we typically develop models of that system so that we can predict how the system will change with time, either autonomously or in response to hypothetical inputs. Hence we can develop multiple models of the same system, where some may be better than others at describing particular features of the system. For instance, there are multiple models used to predict the weather, where one model might be the better at predicting the temperature, while another model be better at predicting whether it will rain.

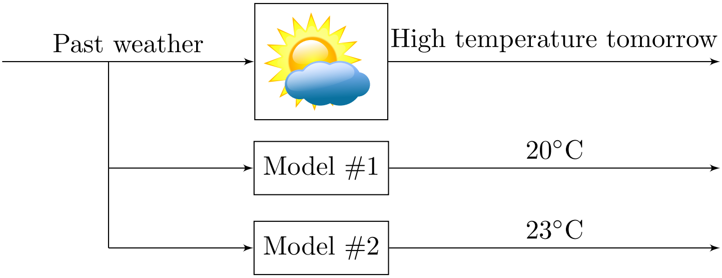


Figure 4: Multiple weather models predicting different temperatures for tomorrow.

A mathematical object, on the other hand, can be referred to as either a *system* or a *model*, depending on the context. If alternative mathematical models are derived from a central mathematical object, then the main object is usually called *the system*, and the derived objects are called *models*. You might want to derive alternative mathematical models if, for instance, the system's mathematical representation is too complicated, or if it does not help you understand the physical principles which drive the system. For example, it is often helpful to linearize a nonlinear system in order to be able to examine its behavior at some point, in which case, we have a linear *model* of a nonlinear *system*.

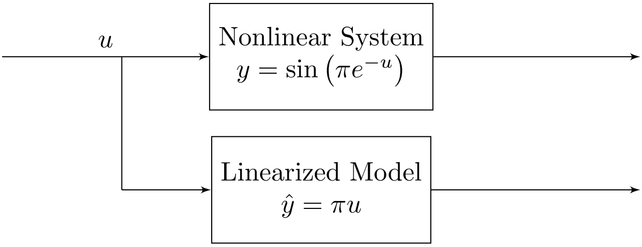


Figure 5: A linearized model of a nonlinear system.

Typically, the first step in the system identification process is to select a model structure for the system that we want to model. For instance, in the second half of Example 1.1, we derived the structure of the model from first principles. We will not always be so fortunate. To the contrary, it will often happen that instead of being given a specific model structure, we only know very macroscopic information about the system. For instance, we might only know that the system's dynamics can be approximately described by a linear ordinary differential equation.

However, in addition to classifying a system as linear or nonlinear, it is also helpful to characterize systems strictly in terms of the types of their inputs u and outputs y, which can be analog or digital, and scalar or multivariable. Furthermore, the inputs and outputs can also belong to an arbitrary field $\mathbb{F}$, although we will almost exclusively consider the case where u and y are real numbers, that is, $u \in \mathbb{R}^m$ and $y \in \mathbb{R}^p$. These distinctions are often given the following names in the literature:

Definition 1.1

- A system is called a *continuous-time* system if u and y are both analog, that is, both u and y are functions of time $t \in \mathbb{R}$.

- A system is called a *discrete-time* system if u and y are both digital, and both sampled at the same frequency, that is, both u and y are functions of a time index $k \in \mathbb{Z}$.

- A system is called a *single-input single-output* (SISO) system if u is scalar ($u \in \mathbb{R}$) and y is scalar ($y \in \mathbb{R}$).

- A system is called a *multiple-input multiple-output* (MIMO) system if u is multivariate ($u \in \mathbb{R}^m$) and y is multivariate ($y \in \mathbb{R}^p$).

□

Remark 1.3 SIMO and MISO systems can be defined analogously, although they will typically be referred to as MIMO systems. □

Example 1.2 Consider the system which relates the force on a car's brake and gas pedals with the car's position and speed, as shown in Figure 6. The car is a MIMO system since there are two inputs and two outputs. Furthermore, if the measurements of the speed, position, and pedal forces are analog, then the system is called a continuous-time system. If the measurements are digital (or sampled from the analog signals), then the system is called a discrete-time system. □

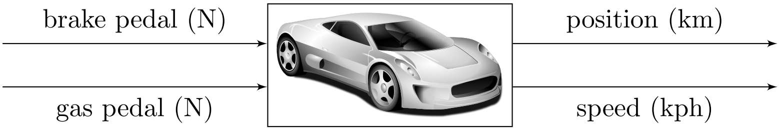


Figure 6: The *system* interpretation of a car, where the brake and gas pedal forces are viewed as inputs, and the position and speed are viewed as outputs.

In the following sections, we show some of the most common models that you will encounter, and describe some properties that a model may have. Furthermore, since the extra effort in dealing with MIMO systems is minimal, we will henceforth always assume that the system is MIMO, unless otherwise noted.

SECTION 2

State-Space Models

State-space models are perhaps the most common type of models considered in the system identification and control literature. Here we present several linear state-space model forms, their solutions, and some of their properties.

Remark 2.1 We use the term *state-space models* to refer to linear model forms, although there also exist *nonlinear state-space models.* We will address linearity more precisely in Section 7. □

2.1 Time-Varying Continuous-Time Models

Time-varying, continuous-time state-space models are models of the form

$$\begin{aligned} \dot{x}(t) &= A(t)x(t) + B(t)u(t) \\ y(t) &= C(t)x(t) + D(t)u(t) \end{aligned} \tag{2.1}$$

where $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, and $y \in \mathbb{R}^p$ are called the *state*, *input*, and *output*, respectively. The state x and output y of (2.1) are equivalently given by the variation of parameters formulas [1, p. 109]:

$$x(t) = \quad \Phi(t,t_0)x(t_0) + \int_{t_0}^{t} \Phi(t,\tau)B(\tau)u(\tau)d\tau \tag{2.2}$$

$$y(t) = C(t)\left[\Phi(t,t_0)x(t_0) + \int_{t_0}^{t} \Phi(t,\tau)B(\tau)u(\tau)d\tau \right] + D(t)u(t) \tag{2.3}$$

where the *state-transition matrix* Φ is a solution of

$$\frac{d\Phi(t,t_0)}{dt} = A(t)\Phi(t,t_0), \quad \text{s.t.} \quad \Phi(t_0,t_0) = I \tag{2.4}$$

Remark 2.2 There may exist more than one solution of (2.4). Hence we say that Φ is *a solution* of (2.4). The following fact gives a sufficient condition for there to exist a unique solution of (2.4) □

Fact 2.1 Consider the state-space model (2.1). If $A(t)$ is continuous, then there exists a unique solution Φ of (2.4). Furthermore, (2.2) and (2.3) are the unique solutions of (2.1).

Proof The first statement is proved in [2, p. 64-65]. The second statement follows by plugging the unique state-transition matrix Φ into (2.2) and (2.3), and verifying that (2.2) and (2.3) are solutions of (2.1). □

Remark 2.3 To show that (2.2) is a solution of (2.1), substitute (2.2) into (2.1), and use Leibniz's integration rule [3, p. 24]:

Fact 2.2 Let $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. If $\partial f(x,y)/\partial x$ is continuous, then

$$\begin{aligned} \frac{\partial}{\partial x}\left[\int\limits_{u(x)}^{v(x)} f(x,y)dy\right] = \int\limits_{u(x)}^{v(x)} \frac{\partial f(x,y)}{\partial x}dy + f\big(x, v(x)\big)\frac{dv(x)}{dx} \\ - f\big(x, u(x)\big)\frac{du(x)}{dx} \end{aligned} \tag{2.5}$$

□

Next, we prove some useful facts about the continuous-time state-transition matrix.

Fact 2.3 Let Φ be the state-transition matrix of the time-varying, continuous-time state-space model (2.1). If $A(t)$ is continuous, then for all times t, t_0, and T, the state-transition matrix $\Phi(t,t_0)$ has full rank. Furthermore,

$$\Phi(t,t_0) = \Big[\Phi(t_0,t)\Big]^{-1} \tag{2.6}$$

$$\Phi(T,t_0) = \Phi(T,t)\Phi(t,t_0) \tag{2.7}$$

Proof First, let the input u to the model (2.1) be zero. Then since $A(t)$ is continuous, from Fact 2.1 and (2.2) we have that the unique solution x of (2.1) is given by

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

Next, let $x_1(t_0), \ldots, x_n(t_0) \in \mathbb{R}^n$ be n linearly independent initial conditions and let

$$X(t) \triangleq \begin{bmatrix} x_1(t) & \cdots & x_n(t) \end{bmatrix}$$

Then integrating (2.1) forward and backward in time between t_0 and t yields

$$X(t) = \Phi(t, t_0)X(t_0) \tag{2.8}$$

$$X(t_0) = \Phi(t_0, t)X(t) \tag{2.9}$$

Hence substituting (2.8) into (2.9), we find that

$$X(t_0) = \Phi(t_0, t)X(t) = \Phi(t_0, t)\Phi(t, t_0)X(t_0)$$

where, since $x_1(t_0), \ldots, x_n(t_0)$ are linearly independent, $X(t_0)$ is invertible. Therefore right-multiplying by $X(t_0)^{-1}$, it follows that

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

Thus $\Phi(t_0, t)$ and $\Phi(t, t_0)$ have full rank and (2.6) holds.

Finally, from (2.8), note that since $X(t_0)$ and $\Phi(t, t_0)$ have full rank for all t, it follows that $X(t)$ has full rank for all t. Hence

$$\Phi(T, t)\Phi(t, t_0) = \Big[X(T)X^{-1}(t)\Big]\Big[X(t)X^{-1}(t_0)\Big] = X(T)X^{-1}(t_0) = \Phi(T, t_0)$$

and therefore (2.7) also holds. □

2.2 Time-Invariant Continuous-Time Models

Time-invariant, continuous-time state-space models are models of the form

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) + Du(t) \end{aligned} \tag{2.10}$$

where $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, and $y \in \mathbb{R}^p$ are called the *state*, *input*, and *output*, respectively. The *state-transition matrix* Φ of (2.10) is given by

$$\Phi(t, t_0) = e^{A[t-t_0]} \tag{2.11}$$

where Φ is the solution of (2.4), and $e^{A[t-t_0]}$ denotes the matrix exponential, that is,

$$e^X = I + X + \frac{X^2}{2!} + \frac{X^3}{3!} + \cdots \tag{2.12}$$

Hence from (2.2)-(2.3), it follows that the state x and output y of (2.10) are given by

$$x(t) = \quad e^{A[t-t_0]}x(t_0) + \int\limits_{t_0}^{t} e^{A[t-\tau]}Bu(\tau)d\tau \tag{2.13}$$

$$y(t) = C\left[e^{A[t-t_0]}x(t_0) + \int\limits_{t_0}^{t} e^{A[t-\tau]}Bu(\tau)d\tau\right] + Du(t) \tag{2.14}$$

Remark 2.4 In (2.1)-(2.4), A is a function of time. In (2.11)-(2.14), $A[t - t_0]$ denotes the product of the constant matrix A with the scalar time $t - t_0$. □

2.3 Time-Varying Discrete-Time Models

Time-varying, discrete-time state-space models are models of the form

$$\begin{aligned} x(k+1) &= A(k)x(k) + B(k)u(k) \\ y(k) &= C(k)x(k) + D(k)u(k) \end{aligned} \tag{2.15}$$

where $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, and $y \in \mathbb{R}^p$ are called the *state*, *input*, and *output*, respectively. Furthermore, if $k \geq k_0$, then the state x and output y of (2.15) are given by [1, p. 111]:

$$x(k) = \qquad \Phi(k,k_0)x(k_0) + \sum_{\kappa=k_0}^{k-1} \Phi(k,\kappa+1)B(\kappa)u(\kappa) \tag{2.16}$$

$$y(k) = C(k)\left[\Phi(k,k_0)x(k_0) + \sum_{\kappa=k_0}^{k-1} \Phi(k,\kappa+1)B(\kappa)u(\kappa)\right] + D(k)u(k) \tag{2.17}$$

where the *state-transition matrix* Φ is given by

$$\Phi(k+1,k_0) = A(k)\Phi(k,k_0) = A(k)\cdots A(k_0) \quad \text{s.t.} \quad \Phi(k_0,k_0) = I \tag{2.18}$$

However, if $k < k_0$ and $A(k),\ldots,A(k_0-1)$ are invertible, then (2.16) also holds, in which case Φ is given by

$$\Phi(k,k_0) = \begin{cases} I, & k = k_0 \\ A(k-1)\cdots A(k_0), & k > k_0 \\ A^{-1}(k)\cdots A^{-1}(k_0-1), & k < k_0 \end{cases} \tag{2.19}$$

Next, we prove some useful facts about the discrete-time state-transition matrix.

Fact 2.4 Let Φ be the state-transition matrix of the time-varying, discrete-time state-space model (2.15). If $K \geq k \geq k_0$, then

$$\Phi(K,k_0) = \Phi(K,k)\Phi(k,k_0) \tag{2.20}$$

Furthermore, if $A(k_0),\ldots,A(K-1)$ are invertible, then $\Phi(K,k_0)$ has full rank, $\Phi(k_0,K)$ has full rank, and

$$\Phi(k_0,K) = \Big[\Phi(K,k_0)\Big]^{-1} \tag{2.21}$$

$$\Phi(k_0,K) = \Phi(k_0,k)\Phi(k,K) \tag{2.22}$$

$$\Phi(k_0,k) = \Phi(k_0,K)\Phi(K,k) \tag{2.23}$$

Proof First, note that (2.20) follows directly from (2.18). Furthermore, if $A(k_0), \ldots, A(K-1)$ are invertible, then

$$\Big[\Phi(K, k_0)\Big]^{-1} = \Big[A(K-1) \cdots A(k_0)\Big]^{-1} = A^{-1}(k_0) \cdots A^{-1}(K-1)$$

Hence comparing with (2.19), we find (2.21). The other two relations follow from (2.19) and (2.20).

Finally, since $A(k_0), \ldots, A(K-1)$ are invertible, then from (2.19), it follows that $\Phi(K, k_0)$ and $\Phi(k_0, K)$ have full rank. □

2.4 Time-Invariant Discrete-Time Models

Time-invariant, discrete-time state-space models are models of the form

$$\begin{aligned} x(k+1) &= Ax(k) + Bu(k) \\ y(k) &= Cx(k) + Du(k) \end{aligned} \tag{2.24}$$

where $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, and $y \in \mathbb{R}^p$ are called the *state*, *input*, and *output*, respectively. If $k \geq k_0$, then the state-transition matrix Φ is given by (2.18), or simply

$$\Phi(k, k_0) = A^{k-k_0} \tag{2.25}$$

Hence if $k \geq k_0$, then from (2.16)-(2.17), the state x and output y are given by

$$x(k) = \quad A^{k-k_0} x(k_0) + \sum_{\kappa=k_0}^{k-1} A^{k-1-\kappa} B u(\kappa) \tag{2.26}$$

$$y(k) = C\left[A^{k-k_0} x(k_0) + \sum_{\kappa=k_0}^{k-1} A^{k-1-\kappa} B u(\kappa)\right] + Du(k) \tag{2.27}$$

Furthermore, if A is invertible, then (2.25)-(2.27) also hold for $k < k_0$.

SECTION 3

Sampling and Zero-Order Holds

Although the dynamics of a system are often modelled by a set of differential equations, such as the continuous-time state-space model

$$\begin{aligned} \dot{x}(t) &= A(t)x(t) + B(t)u(t) \\ y(t) &= C(t)x(t) + D(t)u(t) \end{aligned} \tag{3.1}$$

we will rarely be able to observe or measure the continuous output of such a system. To the contrary, most modern sensors are digital, and therefore bounded by a finite sampling rate. Hence the outputs y of a system are usually only measured at a finite number of equally-spaced times t_0, t_0+h, t_0+2h, etc., as shown in Figure 7, where h is called the *sampling interval*, and $1/h$ is called the *sampling frequency*.

Remark 3.1 You could think of a seismometer, which provides a continuous trace of seismic activity, as one exception to this rule. □

Remark 3.2 The process of measuring a continuous-signal at a finite number of times is called *sampling*, and the measured points are called *samples*. □

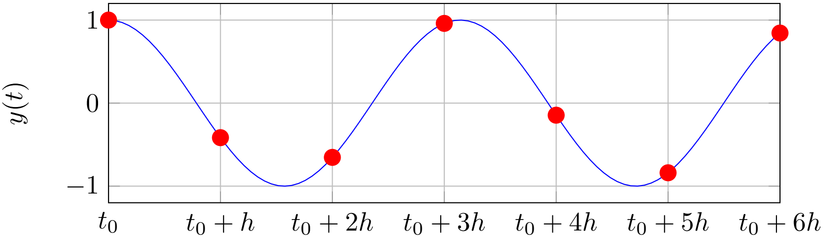


Figure 7: A continuous output signal, where the filled circles denote the sampled values.

Similarly, most modern actuators are digitally controlled, and hence the inputs u can only be changed at a finite number of equally-spaced times t_0, t_0+h, t_0+2h, etc. However, whereas we can't be sure how a sampled signal behaves between the measurement times t_0 and t_0+h, this behavior is usually well-defined for actuators. Specifically, most actuators use the zero-order hold principle, as illustrated in Figure 8, which holds the actuator output constant between input commands.

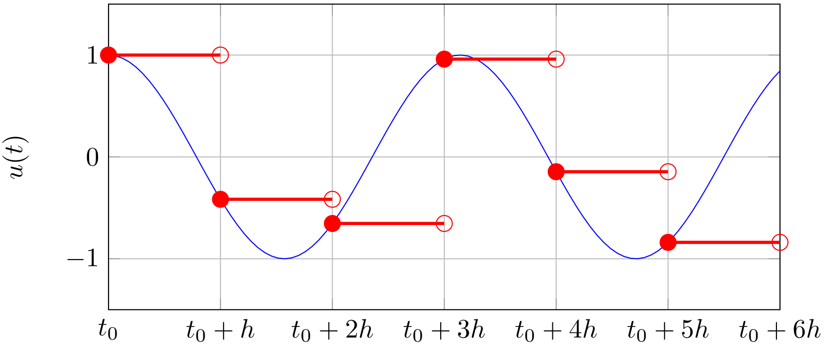


Figure 8: A sinusoidal signal along with the zero-order hold version of the same signal, where the filled circles represent the input commands, and the open circles represent discontinuities. You could think of the blue line as representing the desired input command, and the red line as representing the actual actuator output given to the continuous system.

Remark 3.3 Figure 8 demonstrates a very exaggerated zero-order hold. Usually the sampling rate h is much smaller than the signal period so that the difference between the desired and zero-order hold signals is negligible for engineering purposes. ❒

Remark 3.4 The zero-order hold model is not exactly correct since real actuators are rarely discontinuous. Hence it may sometimes be necessary to construct a more accurate model of the actuator's behavior to achieve the desired model accuracy. ❒

Remark 3.5 There exist many other models for actuator behavior, such as the first-order hold, but the zero-order hold is by far the most common. ❒

3.1 Discretization

The process of converting a continuous-time model into a discrete-time model is called *discretization*. We usually need to compute this transformation because, in a real system, the continuous-time outputs are almost always sampled and the inputs are almost always passed through a zero-order hold. Hence if we want to implement a continuous-time control algorithm on a real system, it will first need to be discretized so that it can be implemented on digital hardware. For this purpose, we will develop an equivalent discrete-time model of (3.1), that is, a model of the form

$$\begin{aligned} x_d(k+1) &= A_d(k)x_d(k) + B_d(k)u(k) \\ y(k) &= C_d(k)x_d(k) + D_d(k)u(k) \end{aligned} \tag{3.2}$$

To accomplish this, recall from the variation of parameters formula (2.2) that

$$x(t_0+h) = \Phi(t_0+h, t_0)x(t_0) + \int_{t_0}^{t_0+h} \Phi(t_0+h, \tau)B(\tau)u(\tau)d\tau \tag{3.3}$$

Hence, if the inputs u are passed through a zero-order hold actuator with the sampling interval h, then

$$x(t_0+h) = \Phi(t_0+h, t_0)x(t_0) + \left[\int_{t_0}^{t_0+h} \Phi(t_0+h, \tau)B(\tau)d\tau\right] u(t_0) \tag{3.4}$$

where the state x and input u are known at times $t_0, t_0+h, t_0+2h, \ldots$ Furthermore, if the output y is sampled at the same interval h, and the measurements are synchronized with the input signal, then for all nonnegative integers i, we have that

$$y(t_0+ih) = C(t_0+ih)x(t_0+ih) + D(t_0+ih)u(t_0+ih) \tag{3.5}$$

Therefore letting

$$\begin{aligned}
x_d(k+i) &\triangleq x(t_0+ih)\\
A_d(k+i) &\triangleq \Phi\big(t_0+[i+1]h, t_0+ih\big)\\
B_d(k+i) &\triangleq \int\limits_{t_0+ih}^{t_0+[i+1]h} \Phi\big(t_0+[i+1]h, \tau\big)B(\tau)d\tau\\
C_d(k+i) &\triangleq C(t_0+ih)\\
D_d(k+i) &\triangleq D(t_0+ih)
\end{aligned}\tag{3.6}$$

we find that the discrete-time model (3.2) is an equivalent representation of (3.1) when the outputs are sampled, and the inputs are passed through a zero-order hold.

Remark 3.6 The state-transition matrix Φ in (3.6) is the continuous-time one, that is, it is the solution of

$$\frac{d\Phi(t,t_0)}{dt} = A(t)\Phi(t,t_0), \quad \text{s.t.} \quad \Phi(t_0,t_0) = I \tag{3.7}$$

□

Remark 3.7 The model (3.6) shows that when the output of a time-varying continuous-time state-space model is sampled, and the input is passed through a zero-order hold actuator with the same sampling interval, then the original model can be interpreted as a time-varying discrete-time model, as in (3.6). However, the reverse is not true, that is, given a time-varying discrete-time model of the form

$$\begin{aligned}
x(k+1) &= A(k)x(k) + B(k)u(k)\\
y(k) &= C(k)x(k) + D(k)u(k)
\end{aligned}\tag{3.8}$$

there does not always exist a continuous-time model from which it was derived. For instance, if $A(k)$ does not have full rank, then there does not exist a continuous-time analog since (3.6) implies that the continuous-time state-transition matrix Φ would also be rank deficient. However, this contradicts Fact 2.3, and hence there is no continuous-time analog for a discrete-time model with a rank-deficient A-matrix. □

SECTION 4

States and Similarity Transformations

State-space models explicitly include a state x, where $x(t_0)$ encapsulates everything that happened in the system before time t_0. Hence given $x(t_0)$, we only need to know the input u for all times $t \geq t_0$ to uniquely determine the output y for all times $t \geq t_0$, as we demonstrated in (2.3), (2.14), (2.17), and (2.27). In fact, this definition of a *state* is useful for all models, even those models which do not explicitly include a state:

Definition 4.1 If the output y of a system is uniquely determined for all times $t \geq t_0$ by

1) the input u for all times $t \geq t_0$
and 2) the information x at time t_0,

then x is called a *state* of the system and $x(t_0)$ is called an *initial condition.* □

Remark 4.1 The *state* refers to the signal x, whereas the *initial condition* is the result of evaluating the state at a specific time, such as t_0. □

It is important to note that Definition 4.1 is actually different to what we have called a state up to this point. For instance, in the state-space model

$$\begin{aligned} \dot{x}(t) &= A(t)x(t) + B(t)u(t) \\ y(t) &= C(t)x(t) + D(t)u(t) \end{aligned} \tag{4.1}$$

we have called the signal $x \in \mathbb{R}^n$ *the state*, as if it is the unique signal from which the output y can be uniquely determined. To the contrary, there are an infinite number of states from which the output y can be uniquely determined, as we demonstrate in the following example.

Example 4.1 Consider the model (4.1). Also, let $P(t) \in \mathbb{R}^{n \times n}$ be invertible for all $t \in \mathbb{R}$, and let

$$x_p(t) \triangleq P^{-1}(t)x(t) \tag{4.2}$$

Then from (4.1), it follows that

$$\begin{aligned} \frac{d}{dt}\Big[P(t)x_p(t)\Big] &= A(t)P(t)x_p(t) + B(t)u(t) \\ y(t) &= C(t)P(t)x_p(t) + D(t)u(t) \end{aligned} \tag{4.3}$$

and hence

$$\begin{aligned} \dot{x}_p(t) &= P^{-1}(t)\Big[A(t)P(t) - \dot{P}(t)\Big]x_p(t) + P^{-1}(t)B(t)u(t) \\ y(t) &= C(t)P(t)x_p(t) + D(t)u(t) \end{aligned} \tag{4.4}$$

Hence the output y of (4.1) is also uniquely determined for all times $t \geq t_0$ by $x_p(t_0)$ and the input u for all times $t \geq t_0$. □

From Definition 4.1 and Example 4.1, we see that there are actually an infinite number of states of the state-space model (4.1), despite the fact that we call the signal x in (4.1) *the state*. The usage of the word *state* is therefore a matter of perspective:

1) When we use a state-space model such as (4.1) with a well-defined A, B, C, and D matrix, then there is clearly a natural state, namely, the one which explicitly appears in the state-space description. Hence we refer to the signal x in the state-space model as *the state*.

2) Similarly, if there is some physical principle at play, then we might refer to some of the underlying physical properties, such as position and velocity, as *the state* of the system, regardless of whether or not we were using a state-space model.

3) However, if we were primarily concerned with the input/output behavior of a system, then the concept of a state is quite arbitrary. For instance, both of the models (4.1) and (4.4) display the same input/output behavior, and hence there is no preference for either model or state. In this case, one must be clear about which state one is referring to.

The class of transformations demonstrated in Example 4.1 are called *similarity transformations*. You will come across then often in the system identification literature. Therefore we formally introduce them in the following definition:

Definition 4.2 Let G and H denote models with the same input u. Furthermore, let $x_g \in \mathbb{R}^{n_g}$ and $x_h \in \mathbb{R}^{n_h}$ denote states of the systems G and H, and let y_g and y_h denote their outputs. Then:

i) The models G and H are *output-equivalent* if both of the following hold:

   a) For all initial conditions $x_g(t_0)$, there exists an initial condition $x_h(t_0)$ such that $y_g(t) = y_h(t)$ for all $t \geq t_0$ and all inputs u.

   b) For all initial conditions $x_h(t_0)$, there exists an initial condition $x_g(t_0)$ such that $y_g(t) = y_h(t)$ for all $t \geq t_0$ and all inputs u.

ii) The models G and H are *equivalent* if they are output-equivalent, and for all $t \in \mathbb{R}$, there exists a one-to-one mapping between the states $x_g(t)$ and $x_h(t)$.

iii) The models G and H are *similar* if they are output-equivalent, and for all $t \in \mathbb{R}$, there exists an invertible $P(t) \in \mathbb{R}^{n_h \times n_g}$ such that

$$x_h(t) = P(t)x_g(t) \tag{4.5}$$

In this case, (4.5) is called a *similarity transformation.*

□

Example 4.1 gives an example of a model which is similar to the the state-space model (4.1). The following example shows that output-equivalence does not imply equivalence or similarity.

Example 4.2 Consider the model (4.1). Also, let

$$x_p(t) \triangleq \begin{bmatrix} I_n \\ I_n \end{bmatrix} x(t) \tag{4.6}$$

Then from (4.1), it follows that

$$\begin{aligned} \dot{x}_p(t) &= \begin{bmatrix} A(t) & 0_{n\times n} \\ 0_{n\times n} & A(t) \end{bmatrix} x_p(t) + \begin{bmatrix} B(t) \\ B(t) \end{bmatrix} u(t) \\ y(t) &= \begin{bmatrix} C(t) & 0_{n\times n} \end{bmatrix} x_p(t) + D(t)u(t) \end{aligned} \tag{4.7}$$

Hence (4.1) and (4.7) are output-equivalent. However, since there exists neither a one-to-one mapping or similarity transform between the states of the systems, (4.1) and (4.7) are neither equivalent or similar. □

4.1 Realizations and Minimal Realizations

TODO

SECTION 5

Controllability and Reachability

Definition 5.1 Let $x : \mathbb{R} \to \mathbb{R}^n$ denote a state of the system G. Then

(i) G is *reachable from* t_0 if, for all initial conditions $x(t_0) \in \mathbb{R}^n$, and all $x_f \in \mathbb{R}^n$, there exists a finite time $t \geq t_0$ and an input $u : [t_0, t] \to \mathbb{R}^m$ such that $x(t) = x_f$.

(ii) G is *completely reachable* if it is reachable from all t_0.

(iii) G is *controllable from* t_0 if, for all initial conditions $x(t_0) \in \mathbb{R}^n$, there exists a finite time $t \geq t_0$ and an input $u : [t_0, t] \to \mathbb{R}^m$ such that $x(t) = 0$.

(iv) G is *completely controllable* if it is controllable from all t_0.

□

Fact 5.1 If a system is reachable from t_0, then it is controllable from t_0. If a system is completely reachable, then it is completely controllable.

Proof Let $x_f = 0$ in the reachable definitions. □

Remark 5.1 The *reachable* and *controllable* definitions in Definition 5.1 are not universal. For instance,

- [4, p. 65] calls a system *controllable* is it satisfies our definition of completely reachable.
- [5, p. 215] defines *reachable* by modifying our definition so that the initial condition is zero, that is, $x(t_0) = 0$.
- [6, p. 94] and [7, p. 64-65] do not use the *completely* distinctions.

Pay careful attention to the definitions used in other books. □

5.1 Time-Varying Continuous-Time Relations

Consider the time-varying, continuous-time state-space model

$$\begin{aligned} \dot{x}(t) &= A(t)x(t) + B(t)u(t) \\ y(t) &= C(t)x(t) + D(t)u(t) \end{aligned} \tag{5.1}$$

Definition 5.2 A *reachability gramian* of the time-varying, continuous-time model (5.1) is of the form

$$W_r(t, t_0) \triangleq \int_{t_0}^{t} \Phi(t, \tau)B(\tau)B^T(\tau)\Phi^T(t, \tau)d\tau \tag{5.2}$$

where the state-transition matrix Φ is a solution of (2.4), that is,

$$\frac{d}{dt}\Phi(t, t_0) = A(t)\Phi(t, t_0), \quad \text{s.t.} \quad \Phi(t_0, t_0) = I \tag{5.3}$$

Furthermore, a *controllability gramian* is of the form

$$W_c(t, t_0) \triangleq \int_{t_0}^{t} \Phi(t_0, \tau)B(\tau)B^T(\tau)\Phi^T(t_0, \tau)d\tau \tag{5.4}$$

❒

Remark 5.2 If $A(t)$ is continuous, then there is a unique solution Φ of (5.3). Hence (5.2) and (5.4) are called *the* reachability and controllabililty gramians. In practice, they are almost always assumed to be unique. ❒

Fact 5.2 Consider the time-varying, continuous-time state-space model (5.1), where Φ denotes the state-transition matrix, and $W_r(t, t_0)$ and $W_c(t, t_0)$ denote the reachability and controllability gramians, respectively. If $A(t)$ is continuous, then

$$W_c(t, t_0) = \Phi(t_0, t)W_r(t, t_0)\Phi^T(t_0, t) \tag{5.5}$$

$$W_r(t, t_0) = \Phi(t, t_0)W_c(t, t_0)\Phi^T(t, t_0) \tag{5.6}$$

Furthermore, $W_r(t, t_0)$ has full rank if and only if $W_c(t, t_0)$ has full rank.

Proof From (5.2) and Fact 2.3, we have that

$$\begin{aligned}\Phi(t_0, t)W_r(t, t_0)\Phi^T(t_0, t) &= \int_{t_0}^{t} \Phi(t_0, t)\Phi(t, \tau)B(\tau)B^T(\tau)\Phi^T(t, \tau)\Phi^T(t_0, t)d\tau \\ &= \int_{t_0}^{t} \Phi(t_0, \tau)B(\tau)B^T(\tau)\Phi^T(t_0, \tau)d\tau = W_c(t, t_0)\end{aligned}$$

Hence (5.5) holds. (5.6) is proved in the same fashion.

Finally, since $A(t)$ is continuous, from Fact 2.3 we find that the state transition matrix Φ always has full rank. Hence from (5.5), it follows that $W_r(t, t_0)$ has full rank if and only if $W_c(t, t_0)$ has full rank. ❒

Fact 5.3 Consider the time-varying, continuous-time model (5.1). If $A(t)$ is continuous, then the following statements are equivalent:

(i) For all t_0, there exists a finite $t \geq t_0$ for which the reachability gramian (5.2) has full rank.

(ii) The continuous-time model (5.1) is completely reachable.

(iii) The continuous-time model (5.1) is completely controllable.

(iv) For all t_0, there exists a finite $t \geq t_0$ for which the controllability gramian (5.4) has full rank.

Specifically, if $x(t_0)$ denotes the initial condition, and x_f denotes the desired final state, then using the input

$$u(\tau) = B^T(\tau)\Phi^T(t,\tau)W_r^{-1}(t,t_0)\Big[x_f - \Phi(t,t_0)x(t_0)\Big] \tag{5.7}$$

for $\tau \in [t_0, t]$ yields the state $x(t) = x_f$. Furthermore, using the input

$$u(\tau) = -B^T(\tau)\Phi^T(t_0,\tau)W_c^{-1}(t,t_0)x(t_0) \tag{5.8}$$

for $\tau \in [t_0, t]$ yields the state $x(t) = 0$.

Proof We will show that (i) ⇒ (ii) ⇒ (iii) ⇒ (iv) ⇒ (i).

(i) ⇒ (ii): Since $A(t)$ is continuous, then from Fact 2.1, there exists a unique solution x of (5.1). Furthermore, letting t denote the time at which the reachability gramian $W_r(t,t_0)$ has full rank, and using the input (5.7), we find that the unique state $x(t)$ of (5.1) is given by

$$\begin{aligned}
x(t) &= \Phi(t,t_0)x(t_0) \\
&\quad + \int_{t_0}^{t} \left(\Phi(t,\tau)B(\tau)B^T(\tau)\Phi^T(t,\tau)W_r^{-1}(t,t_0)\Big[x_f - \Phi(t,t_0)x(t_0)\Big]\right) d\tau \\
&= \Phi(t,t_0)x(t_0) \\
&\quad + \left[\int_{t_0}^{t} \Phi(t,\tau)B(\tau)B^T(\tau)\Phi^T(t,\tau)d\tau\right] W_r^{-1}(t,t_0)\Big[x_f - \Phi(t,t_0)x(t_0)\Big] \\
&= \Phi(t,t_0)x(t_0) + \Big[W_r(t,t_0)\Big]W_r^{-1}(t,t_0)\Big[x_f - \Phi(t,t_0)x(t_0)\Big] = x_f
\end{aligned}$$

(ii) ⇒ (iii): Fact 5.1.

(iii) ⇒ (iv): Let $x_1(t_0), \ldots, x_n(t_0)$ denote n linearly independent initial conditions. Then since the system is completely controllable, for each initial condition $x_i(t_0)$ there exists a finite time t_i and an input $u_i : [t_0, t_i] \to \mathbb{R}^m$ such that

$$x(t_i) = 0 = \Phi(t_i, t_0)x_i(t_0) + \int_{t_0}^{t_i} \Phi(t_i, \tau)B(\tau)u_i(\tau)d\tau \tag{5.9}$$

Next, let $T \triangleq \max(t_1, \ldots, t_n)$, $X(t_0) \triangleq \left[\begin{array}{ccc} x_1(t_0) & \cdots & x_n(t_0) \end{array}\right]$, and

$$\tilde{u}_i(\tau) \triangleq \begin{cases} u_i(\tau), & \tau \in [t_0, t_i] \\ 0, & \text{otherwise} \end{cases}$$

Then left-multiplying (5.9) by $\Phi(t_0, T)$, and using Fact 2.3, we find that

$$X(t_0) = -\int_{t_0}^{T} \Phi(t_0, \tau)B(\tau) \left[\begin{array}{ccc} \tilde{u}_1(\tau) & \cdots & \tilde{u}_n(\tau) \end{array}\right] d\tau \tag{5.10}$$

Finally, suppose that (iv) does not hold, that is, suppose that there exists a initial t_0 from which the controllability gramian (5.4) is singular for all finite $t \geq t_0$. Then for all finite $t \geq t_0$, there exists a nonzero vector $v(t)$ in the nullspace of $W_c(t, t_0)$, that is,

$$\begin{aligned} v^T(t)W_c(t, t_0)v(t) &= \int_{t_0}^{t} v^T(t)\Phi(t_0, \tau)B(\tau)B^T(\tau)\Phi^T(t_0, \tau)v(t)d\tau \\ &= \int_{t_0}^{t} \left\| v^T(t)\Phi(t_0, \tau)B(\tau) \right\|_2^2 d\tau = 0 \end{aligned}$$

Specifically, since the norm of a vector is nonnegative, we find that for all finite $t \geq t_0$ and all $\tau \in [t_0, t]$, there exists a nonzero vector $v(t)$ such that

$$v^T(t)\Phi(t_0, \tau)B(\tau) = 0$$

Hence left-multiplying (5.10) by $v^T(T)$, we find that

$$v^T(T)X(t_0) = -v^T(T) \int_{t_0}^{T} \Phi(t_0, \tau)B(\tau) \left[\begin{array}{ccc} \tilde{u}_1(\tau) & \cdots & \tilde{u}_n(\tau) \end{array}\right] d\tau = 0$$

However, since $v(T)$ is nonzero, this contradicts the fact that the initial conditions $x_1(t_0), \ldots, x_n(t_0)$ are linearly independent. Hence (iv) holds.

(iv) ⇒ (i): Fact 5.2. □

Fact 5.4 Consider the time-varying, continuous-time model (5.1). If $A(t)$ is continuous, then the following statements are equivalent:

(i) There exists a finite $t \geq t_0$ for which the reachability gramian (5.2) has full rank.

(ii) The continuous-time model (5.1) is reachable from t_0.

(iii) The continuous-time model (5.1) is controllable from t_0.

(iv) There exists a finite $t \geq t_0$ for which the controllability gramian (5.4) has full rank.

Proof This proof is a special case of the proof of Fact 5.3. □

5.2 Time-Varying Discrete-Time Relations

Consider the time-varying, discrete-time state-space model

$$\begin{aligned} x(k+1) &= A(k)x(k) + B(k)u(k) \\ y(k) &= C(k)x(k) + D(k)u(k) \end{aligned} \tag{5.11}$$

Definition 5.3 The *reachability gramian* of the time-varying discrete-time model (5.11) is given by

$$W_r(k, k_0) \triangleq \sum_{\kappa=k_0}^{k-1} \Phi(k, \kappa+1)B(\kappa)B^T(\kappa)\Phi^T(k, \kappa+1), \qquad k > k_0 \tag{5.12}$$

where $W_r(k_0, k_0) = 0$, and the state-transition matrix Φ is given by (2.18), that is,

$$\Phi(k+1, k_0) = A(k)\Phi(k, k_0) = A(k) \cdots A(k_0) \quad \text{s.t.} \quad \Phi(k_0, k_0) = I \tag{5.13}$$

□

Fact 5.5 Consider the time-varying, discrete-time model (5.11). The following statements are equivalent:

(i) For all k_0, there exists a finite $k \geq k_0$ for which the reachability gramian (5.12) has full rank.

(ii) The discrete-time model (5.11) is completely reachable.

Specifically, if $x(k_0)$ denotes the initial condition, and x_f denotes the desired final state, then using the input

$$u(\kappa) = B^T(\kappa)\Phi^T(k, \kappa+1)W_r^{-1}(k, k_0)\Big[x_f - \Phi(k, k_0)x(k_0)\Big] \tag{5.14}$$

for $\kappa \in [k_0, k]$ yields the state $x(k) = x_f$.

Proof (i) ⇒ (ii): Let k denote the time at which $W_r(k, k_0)$ has full rank. Then using the input (5.14), we find that the state x of (5.11) is given by

$$x(k) = \Phi(k, k_0)x(k_0) + W_r(k, k_0)W_r^{-1}(k, k_0)\Big[x_f - \Phi(k, k_0)x(k_0)\Big] = x_f$$

(ii) ⇒ (i): Let $x(k_0) = 0$ and

$$\begin{aligned} \mathcal{C}(k, k_0) &\triangleq \begin{bmatrix} \Phi(k, k)B(k-1) & \cdots & \Phi(k, k_0{+}1)B(k_0) \end{bmatrix} \\ \mathcal{U}(k, k_0) &\triangleq \begin{bmatrix} u^T(k-1) & \cdots & u^T(k_0) \end{bmatrix}^T \end{aligned}$$

Then the state x of (5.11) is given by

$$x(k) = \Phi(k, k_0)x(k_0) + \mathcal{C}(k, k_0)\mathcal{U}(k, k_0) = \mathcal{C}(k, k_0)\mathcal{U}(k, k_0)$$

Next, let $x_1, \ldots, x_n \in \mathbb{R}^n$ denote n linearly independent choices of x_f. Then since the system is completely reachable, for each choice x_i of x_f, there exists a finite time k_i and an input u such that $x(k_i) = x_i$, that is,

$$\begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix} = \begin{bmatrix} \mathcal{C}(k_1, k_0)\mathcal{U}_1(k_1, k_0) & \cdots & \mathcal{C}(k_n, k_0)\mathcal{U}_n(k_n, k_0) \end{bmatrix}$$

where $\mathcal{U}_i(k_i, k_0)$ denotes the matrix $\mathcal{U}$ constructed from the input u necessary to guide the system from the initial condition $x(k_0) = 0$ to the desired final state $x(k_i) = x_i$. Hence letting $K \triangleq \max(k_1, \ldots, k_n)$, we find that

$$\begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix} = \mathcal{C}(K, k_0)\begin{bmatrix} 0_{(K-k_1)m} & \cdots & 0_{(K-k_n)m} \\ \mathcal{U}_1(k_1, k_0) & \cdots & \mathcal{U}_n(k_n, k_0) \end{bmatrix} \tag{5.15}$$

where m denotes the dimension of $u(k)$, that is, $u(k) \in \mathbb{R}^m$.

Finally, since $x_1, \ldots, x_n$ are linearly independent, it follows that the left hand side of (5.15) has full rank, and therefore $\mathcal{C}(K, k_0)$ must have full row rank. Furthermore, since $\mathcal{C}(K, k_0)$ must have full row rank, $\mathcal{C}(K, k_0)\mathcal{C}^T(K, k_0)$ must have full rank, where

$$\begin{aligned}\mathcal{C}(K, k_0)\mathcal{C}^T(K, k_0) &= \sum_{\kappa=k_0}^{K-1} \Phi(K, \kappa+1)B(\kappa)B^T(\kappa)\Phi^T(K, \kappa+1) \\ &= W_r(K, k_0)\end{aligned}$$

Hence the reachability gramian must have full rank at time K. □

Fact 5.6 Consider the time-varying, discrete-time state-space model (5.11). The following statements are equivalent:

(i) There exists a finite $k \geq k_0$ for which the reachability gramian (5.12) has full rank.

(ii) The discrete-time model (5.11) is reachable from k_0.

Proof This proof is a special case of the proof of Fact 5.5. □

5.2.1 Discrete-Time Controllability

A large source of confusion between reachability and controllability is due to the fact that they are equivalent in the continuous-time domain (Facts 5.3 and 5.4), but not in the discrete-time case, as we show with the following example.

Example 5.1 Consider the time-varying, discrete-time, state-space model (5.11), where

$$A(k) = 0, \qquad B(k) = 0, \qquad \text{for all } k$$

Then for all k_0 and all initial conditions $x(k_0)$, there exists a time k and an input u such that $x(k) = 0$. Specifically, at $k = k_0 + 1$, we find that $x(k) = 0$ regardless of the input. Hence the model is completely controllable.

However, computing the reachability gramian, we find that for all k_0 and all $k \geq k_0$, $W_r(k, k_0) = 0$. Hence from Fact 5.5, the model is not completely reachable. □

Example 5.1 shows that controllability does not imply reachability in the discrete-time case. However, there are even more complications in the discrete-time case; if we try to define a discrete-time controllability gramian in the same way as the continuous-time version (5.4), then we need to impose the constraint that A is invertible so that the backward state-transition matrix is defined (Fact 2.4). Specifically, note that the continuous-time controllability gramian includes the state-transition matrix $\Phi(t_0, \tau)$, which, when the input u is 0, satisfies

$$\text{Continuous-time:} \qquad x(t_0) = \Phi(t_0, \tau)x(\tau), \qquad \tau \geq t_0 \tag{5.16}$$

that is, $\Phi(t_0, \tau)$ has the effect of integrating the state backward in time from $\tau \geq t_0$ to t_0, yielding the exact initial condition $x(t_0)$ which resulted in the state $x(\tau)$.

In the discrete-time case, if $k \geq k_0$, then there does not exist a backward state-transition matrix $\Phi(k_0, k)$ unless $A(k_0), \ldots, A(k-1)$ are invertible. For instance, Example 5.1 demonstrates a case where $\Phi(k_0, k)$ does not exist, since it is impossible to determine which initial condition $x(k_0)$ resulted in the state $x(k) = 0$. Specifically, all initial conditions $x(k_0)$ will result in $x(k) = 0$, and hence there does not exist a Φ satisfying

$$x(k_0) = \Phi(k_0, k)x(k), \qquad \text{for all } k \geq k_0 \text{ and all } x(k_0) \tag{5.17}$$

Definition 5.4 Consider the time-varying, discrete-time state-space model (5.11), where $k > k_0$ and $A(k_0), \ldots, A(k-1)$ are invertible. Then the *controllability gramian* is given by

$$W_c(k, k_0) \triangleq \sum_{\kappa=k_0}^{k-1} \Phi(k_0, \kappa+1)B(\kappa)B^T(\kappa)\Phi^T(k_0, \kappa+1), \qquad k > k_0 \tag{5.18}$$

where $W_c(k_0, k_0) = 0$ and the state-transition matrix Φ is given by (2.19), that is,

$$\Phi(k, k_0) = \begin{cases} I, & k = k_0 \\ A(k-1)\cdots A(k_0), & k > k_0 \\ A^{-1}(k)\cdots A^{-1}(k_0-1), & k < k_0 \end{cases} \tag{5.19}$$

□

Fact 5.7 Consider the time-varying, discrete-time state-space model (5.11), where Φ denotes the state-transition matrix and $W_r(k, k_0)$ and

$W_c(k, k_0)$ denote the reachability and controllability gramians, respectively. If $k > k_0$ and $A(k_0), \ldots, A(k-1)$ are invertible, then

$$W_c(k, k_0) = \Phi(k_0, k) W_r(k, k_0) \Phi^T(k_0, k) \tag{5.20}$$

$$W_r(k, k_0) = \Phi(k, k_0) W_c(k, k_0) \Phi^T(k, k_0) \tag{5.21}$$

Furthermore, $W_r(k, k_0)$ has full rank if and only if $W_c(k, k_0)$ has full rank.

Proof From (5.12) and Fact 2.4, we have that

$$\begin{aligned}
&\Phi(k_0, k) W_r(k, k_0) \Phi^T(k_0, k) \\
&\qquad \triangleq \Phi(k_0, k) \sum_{\kappa=k_0}^{k-1} \Phi(k, \kappa+1) B(\kappa) B^T(\kappa) \Phi^T(k, \kappa+1) \Phi^T(k_0, k) \\
&\qquad \triangleq \sum_{\kappa=k_0}^{k-1} \Phi(k_0, k) \Phi(k, \kappa+1) B(\kappa) B^T(\kappa) \Phi^T(k, \kappa+1) \Phi^T(k_0, k) \\
&\qquad \triangleq \sum_{\kappa=k_0}^{k-1} \Phi(k_0, \kappa+1) B(\kappa) B^T(\kappa) \Phi^T(k_0, \kappa+1) \\
&\qquad = W_c(k, k_0)
\end{aligned}$$

Hence (5.20) holds. (5.21) is proved in the same fashion.

Finally, since $A(k_0), \ldots, A(k-1)$ are invertible, then from Fact 2.4 we find that the state-transition matrices $\Phi(k_0, k)$ and $\Phi(k, k_0)$ have full rank. Hence from (5.20), it follows that $W_r(k, k_0)$ has full rank if and only if $W_c(k, k_0)$ has full rank. □

Fact 5.8 Consider the time-varying, discrete-time, state-space model (5.11). If $A(k)$ is invertible for all k, then the following statements are equivalent:

(i) For all k_0, there exists a finite $k \geq k_0$ for which the reachability gramian (5.12) has full rank.

(ii) The continuous-time model (5.11) is completely reachable.

(iii) The continuous-time model (5.11) is completely controllable.

(iv) For all k_0, there exists a finite $k \geq k_0$ for which the controllability gramian (5.18) has full rank.

Specifically, if $x(k_0)$ denotes the initial condition, then the input

$$u(\kappa) = -B^T(\kappa)\Phi^T(k_0, \kappa+1)W_c^{-1}(k, k_0)x(k_0) \tag{5.22}$$

yields the state $x(k) = 0$.

Proof We will show that (i) ⇒ (ii) ⇒ (iii) ⇒ (iv) ⇒ (i).

(i) ⇒ (ii): Fact 5.5.

(ii) ⇒ (iii): Fact 5.1.

(iii) ⇒ (iv): Let

$$\begin{aligned} \tilde{\mathcal{C}}(k, k_0) &\triangleq \begin{bmatrix} \Phi(k_0, k)B(k-1) & \cdots & \Phi(k_0, k_0+1)B(k_0) \end{bmatrix} \\ \mathcal{U}(k, k_0) &\triangleq \begin{bmatrix} u^T(k-1) & \cdots & u^T(k_0) \end{bmatrix}^T \end{aligned}$$

Then the state x of (5.11) is given by

$$\begin{aligned} x(k) &= \Phi(k, k_0)x(k_0) + \Phi(k, k_0)\tilde{\mathcal{C}}(k, k_0)\mathcal{U}(k, k_0) \\ &= \Phi(k, k_0)\Big(x(k_0) + \tilde{\mathcal{C}}(k, k_0)\mathcal{U}(k, k_0)\Big) \end{aligned}$$

Next, let $x_1, \ldots, x_n \in \mathbb{R}^n$ denote n linearly independent choices of $x(k_0)$, and let

$$X \triangleq \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}$$

Then since the system is completely controllable, for each choice x_i of $x(k_0)$, there exists a finite time k_i and an input u such that $x(k_i) = 0$, that is,

$$0 = \Phi(k, k_0)\Bigg(X + \begin{bmatrix} \tilde{\mathcal{C}}(k_1, k_0)\mathcal{U}_1(k_1, k_0) & \cdots & \tilde{\mathcal{C}}(k_n, k_0)\mathcal{U}_n(k_n, k_0) \end{bmatrix}\Bigg)$$

where, since $A(k_0), \ldots, A(k-1)$ are invertible, then from (5.19), $\Phi(k, k_0)$ has full rank. Hence

$$X = -\begin{bmatrix} \tilde{\mathcal{C}}(k_1, k_0)\mathcal{U}_1(k_1, k_0) & \cdots & \tilde{\mathcal{C}}(k_n, k_0)\mathcal{U}_n(k_n, k_0) \end{bmatrix}$$

where $\mathcal{U}_i(k_i, k_0)$ denotes the matrix $\mathcal{U}$ constructed from the input u necessary to guide the system from the initial condition $x(k_0) = x_i$ to $x(k_i) = 0$. Therefore, letting $K \triangleq \max(k_1, \ldots, k_n)$, we find that

$$X = -\tilde{\mathcal{C}}(K, k_0)\begin{bmatrix} 0_{(K-k_1)m} & \cdots & 0_{(K-k_n)m} \\ \mathcal{U}_1(k_1, k_0) & \cdots & \mathcal{U}_n(k_n, k_0) \end{bmatrix} \tag{5.23}$$

where m denotes the dimension of $u(k)$, that is, $u(k) \in \mathbb{R}^m$.

Finally, since $x_1, \ldots, x_n$ are linearly independent, it follows that X has full rank, and therefore $\tilde{\mathcal{C}}(K, k_0)$ has full row rank. Furthermore, since $\tilde{\mathcal{C}}(K, k_0)$ has full row rank, $\tilde{\mathcal{C}}(K, k_0)\tilde{\mathcal{C}}^T(K, k_0)$ has full rank, where

$$\begin{aligned}
\tilde{\mathcal{C}}(K, k_0)\tilde{\mathcal{C}}^T(K, k_0) &= \sum_{\kappa=k_0}^{K-1} \Phi(k_0, \kappa+1)B(\kappa)B^T(\kappa)\Phi^T(k_0, \kappa+1) \\
&= W_c(K, k_0)
\end{aligned}$$

Hence the controllability gramian has full rank at time K.

(iv) ⇒ (i): Fact 5.7.

Finally, we show that (5.22) accomplishes its goal. Specifically, letting k denote the time at which $W_c(k, k_0)$ has full rank, and using the input (5.22), we find that the unique state x of (5.11) is given by

$$\begin{aligned}
x(k) &= \Phi(k, k_0)x(k_0) + \sum_{\kappa=k_0}^{k-1} \Phi(k, \kappa+1)B(\kappa)u(\kappa) \\
&= \Phi(k, k_0)x(k_0) - \Phi(k, k_0)W_c(k, k_0)W_c^{-1}(k, k_0)x(k_0) = 0
\end{aligned}$$

□

5.2.2 Discrete-Time Controllability Warning

You should almost never use the discrete-time controllability gramian given in Definition 5.18 since it requires the invertibility of the A matrices. Specifically, you can get the much stronger result of reachability without requiring invertibility (Facts 5.5 and 5.6), in which case you can use the input (5.14) to achieve whichever final state x_f you want, including $x_f = 0$.

5.3 Time-Invariant Relations

In the time-invariant, continuous-time case, the state-space model simplifies to

$$\begin{aligned}
\dot{x}(t) &= Ax(t) + Bu(t) \\
y(t) &= Cx(t) + Du(t)
\end{aligned} \tag{5.24}$$

and the reachability and controllability gramians simplify to

$$W_r(t,t_0) \triangleq \int_{t_0}^{t} e^{A[t-\tau]} B B^T \left(e^{A[t-\tau]}\right)^T d\tau \tag{5.25}$$

$$W_c(t,t_0) \triangleq \int_{t_0}^{t} e^{A[t_0-\tau]} B B^T \left(e^{A[t_0-\tau]}\right)^T d\tau \tag{5.26}$$

Similarly, for $k > k_0$, the time-invariant, discrete-time model becomes

$$\begin{aligned} x(k+1) &= Ax(k) + Bu(k) \\ y(k) &= Cx(k) + Du(k) \end{aligned} \tag{5.27}$$

and reachability and controllability gramians simplify to

$$W_r(k,k_0) \triangleq \sum_{\kappa=0}^{k-k_0-1} A^{\kappa} B B^T \left(A^{\kappa}\right)^T \tag{5.28}$$

$$W_c(k,k_0) \triangleq \sum_{\kappa=1}^{k-k_0} A^{-\kappa} B B^T \left(A^{-\kappa}\right)^T \tag{5.29}$$

where $W_r(k_0,k_0) = W_c(k_0,k_0) = 0$, and $W_c(k,k_0)$ is only defined if A is invertible.

Although the time-invariant gramians are simpler than their time-varying counterparts, they are still much more complicated than they need to be for the simple task of checking controllability and reachability. Instead, we can simply check the rank of the *controllability matrix*:

Definition 5.5 The *controllability matrix* of a time-invariant, continuous or discrete-time state-space model, such as (5.24) or (5.27), is given by

$$\mathcal{C} \triangleq \begin{bmatrix} B & AB & A^2B & \cdots & A^{n-1}B \end{bmatrix} \tag{5.30}$$

where n is the dimension of the matrix A, that is, $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, and $\mathcal{C} \in \mathbb{R}^{n\times nm}$. □

We will find the controllability matrix particularly useful because of the Cayley-Hamiton theorem [8, p. 265]:

Theorem 5.1 Let $X \in \mathbb{R}^{n\times n}$, and let $p(s)$ denote the characteristic polynomial of X, that is,

$$p(s) = \texttt{det}\Big[sI - X\Big] = p_0 + p_1 s + \cdots + p_{n-1}s^{n-1} + s^n \tag{5.31}$$

Then

$$p(X) = p_0 I + p_1 X + \cdots + p_{n-1}X^{n-1} + X^n = 0 \tag{5.32}$$

An immediate consequence of the Cayley-Hamiton theorem is the following:

Corollary 5.1 Let $X \in \mathbb{R}^{n\times n}$. Then for all $i \geq n$, there exist $\tilde{p}_{i,0}$, …, $\tilde{p}_{i,n-1}$ such that

$$X^i = \tilde{p}_{i,0}I + \tilde{p}_{i,1}X + \cdots + \tilde{p}_{i,n-1}X^{n-1} \tag{5.33}$$

Furthermore, there exist $\tilde{\tilde{p}}_0, \ldots, \tilde{\tilde{p}}_{n-1}$ such that

$$e^X = \tilde{\tilde{p}}_0 I + \tilde{\tilde{p}}_1 X + \cdots + \tilde{\tilde{p}}_{n-1}X^{n-1} \tag{5.34}$$

where e^X denotes the matrix exponential (2.12).

Proof Let $i = n$. Then from the Cayley-Hamilton theorem, specifically (5.32), we have that

$$X^n = -p_0 I - p_1 X - \cdots - p_{n-1}X^{n-1} \tag{5.35}$$

Hence (5.33) holds for $i = n$. Furthermore, multiplying (5.35) by X, it follows that

$$X^{n+1} = -p_0 X - p_1 X^2 - \cdots - p_{n-1}X^n \tag{5.36}$$

Therefore substituting (5.35) into (5.36) for X^n, we find that

$$X^{n+1} = -p_0 X - p_1 X^2 - \cdots - p_{n-1}\Big(- p_0 I - p_1 X - \cdots - p_{n-1}X^{n-1}\Big)$$

Thus (5.33) holds for $i = n+1$, and continuing with this procedure, we find that (5.33) holds for all $i \geq n$.

Finally, recall that $e^X = \sum_{i=0}^{\infty} X^i/i!$. Hence from (5.33) it follows that

$$\begin{aligned} e^X &= I + X + \frac{X^2}{2!} + \cdots + \frac{X^{n-1}}{(n-1)!} \\ &\quad + \sum_{i=n}^{\infty} \frac{1}{i!}\Big(\tilde{p}_{i,0}I + \tilde{p}_{i,1}X + \cdots + \tilde{p}_{i,n-1}X^{n-1}\Big) \end{aligned}$$

□

The final piece of the puzzle is the fact that the matrix exponential always has full rank:

Fact 5.9 Let $X \in \mathbb{C}^{n\times n}$. Then the matrix exponential e^X given by (2.12) has full rank.

Proof See [8, p. 713]. □

Fact 5.10 The following statements are equivalent:

(i) The time-invariant continuous-time model (5.24) is completely reachable.

(ii) The time-invariant continuous-time model (5.24) is completely controllable.

(iii) The controllability matrix $\mathcal{C}$ given in (5.30) has full row rank.

Proof The equivalence of (i) and (ii) was already proved in Fact 5.3. Hence we only need to show that (ii) ⇒ (iii) and (iii) ⇒ (ii).

(ii) ⇒ (iii): From Fact 5.3, if (5.24) is completely controllable, then for all t_0, there exists a finite $t \geq t_0$ for which the reachability gramian has full rank. Furthermore, from Corollary 5.1 we have that

$$e^{A(t-\tau)} = p_0(t-\tau)I + p_1(t-\tau)A + \cdots + p_{n-1}(t-\tau)A^{n-1}$$

Thus letting

$$\mathcal{P}(t-\tau) \triangleq \begin{bmatrix} p_0(t-\tau)I & \cdots & p_{n-1}(t-\tau)I \end{bmatrix}^T$$

it follows that $e^{A(t-\tau)}B = \mathcal{C}\mathcal{P}(t-\tau)$, and therefore

$$W_r(t,t_0) = \mathcal{C}\left(\int_{t_0}^{t} \mathcal{P}(t-\tau)\mathcal{P}^T(t-\tau)d\tau\right)\mathcal{C}^T$$

Finally, since there exists a finite $t \geq t_0$ for which the reachability gramian has full rank, the controllability matrix must have full row rank.

(iii) ⇒ (ii): Suppose that the controllability matrix has full rank, but the system is not completely controllable. Then from Fact 5.3, there exists a t_0 for which the reachability gramian is singular for all finite $t \geq t_0$.

Specifically, referring to (**??**) in the proof of Fact 5.3, for all finite times $t \geq t_0$, there exists a nonzero $v(t)$ satisfying

$$v^T(t)e^{A(t_0-\tau)}B = 0 \quad \text{for all finite } t \geq t_0 \text{ and all } \tau \in [t_0, t]$$

that is, there exists a constant nonzero vector $v_\infty = \lim_{s\to\infty} v(s)$ such that

$$v_\infty^T e^{A(t_0-\tau)}B = 0 \quad \text{for all finite } \tau \geq t_0 \tag{5.37}$$

Hence differentiating (5.37) i times with respect to τ, we find that

$$v_\infty^T e^{A(t_0-\tau)}A^i B = 0$$

and therefore

$$v_\infty^T e^{A(t_0-\tau)}\mathcal{C} = 0$$

However, since the matrix exponential has full rank (Fact 5.9) and $\mathcal{C}$ has full row rank, it follows that $v_\infty^T = 0$, which contradicts the fact that there exists a nonzero v_∞ satisfying (5.37). □

Fact 5.11 The following statements are equivalent:

(i) The time-invariant discrete-time model (5.27) is completely reachable.

(ii) The controllability matrix $\mathcal{C}$ given in (5.30) has full row rank.

Specifically, if $x(k_0)$ denotes the initial condition, and x_f denotes the desired final state, then the input

$$u(k_0+\kappa) = B^T A^{n-1-\kappa}\left(\mathcal{C}\mathcal{C}^T\right)^{-1}\left(x_f - A^n x(k_0)\right) \tag{5.38}$$

yields the state $x(k_0+n) = x_f$. Alternatively, the inputs are given in matrix form by

$$\begin{bmatrix} u(k_0+n-1) \\ \vdots \\ u(k_0) \end{bmatrix} = \mathcal{C}^T\left(\mathcal{C}\mathcal{C}^T\right)^{-1}\left(x_f - A^n x(k_0)\right) \tag{5.39}$$

Proof From Corollary 5.1, recall that for all $\kappa \geq n$, there exists $\tilde{p}_{\kappa,0}$, ..., $\tilde{p}_{\kappa,n-1}$ such that

$$A^\kappa B = \mathcal{C}\left[\begin{array}{ccc} \tilde{p}_{\kappa,0}I & \cdots & \tilde{p}_{\kappa,n-1}I \end{array}\right]^T = \mathcal{C}\tilde{\mathcal{P}}_\kappa \tag{5.40}$$

Therefore, if $k - k_0 \geq n$, then the discrete-time reachability gramian is of the form

$$\begin{aligned} W_r(k,k_0) &= \sum_{\kappa=0}^{n-1} A^{\kappa} B B^T \left[A^{\kappa}\right]^T + \sum_{\kappa=n}^{k-k_0-1} A^{\kappa} B B^T \left[A^{\kappa}\right]^T \\ &= \mathcal{C}\mathcal{C}^T + \sum_{\kappa=n}^{k-k_0-1} \mathcal{C}\tilde{\mathcal{P}}_{\kappa}\tilde{\mathcal{P}}_{\kappa}^T\mathcal{C}^T = \mathcal{C}\left(I + \sum_{\kappa=n}^{k-k_0-1} \tilde{\mathcal{P}}_{\kappa}\tilde{\mathcal{P}}_{\kappa}^T\right)\mathcal{C}^T \end{aligned} \tag{5.41}$$

Furthermore, since $\tilde{\mathcal{P}}_{\kappa}\tilde{\mathcal{P}}_{\kappa}^T$ is positive semi-definite, $I + \sum\limits_{\kappa=n}^{k-k_0-1} \tilde{\mathcal{P}}_{\kappa}\tilde{\mathcal{P}}_{\kappa}^T$ is positive definite, that is, it has full rank. Hence for all k_0 and all $k \geq k_0+n$, $W_r(k,k_0)$ has full rank if and only if $\mathcal{C}$ has full row rank. Therefore, from Fact 5.5, we have that (i) and (ii) are equivalent.

Finally, if the controllability matrix is invertible, then using the input (5.38), we find that the state $x(k_0+n)$ of (5.27) is given by

$$\begin{aligned} x(k_0+n) &= A^n x(k_0) + \sum_{\kappa=k_0}^{k_0+n-1} A^{k_0+n-1-\kappa} B u(\kappa) \\ &= A^n x(k_0) + \sum_{\kappa=0}^{n-1} A^{n-1-\kappa} B u(k_0+\kappa) \\ &= A^n x(k_0) + \sum_{\kappa=0}^{n-1} A^{n-1-\kappa} B B^T A^{n-1-\kappa} \left(\mathcal{C}\mathcal{C}^T\right)^{-1} \left[x_f - A^n x(k_0)\right] \\ &= A^n x(k_0) + \left(\mathcal{C}\mathcal{C}^T\right)\left(\mathcal{C}\mathcal{C}^T\right)^{-1}\left[x_f - A^n x(k_0)\right] = x_f \end{aligned}$$

□

SECTION 6

Observability

Definition 6.1 Let x denote the state of the system G. Then

(i) G is *observable from* t_0 if, for all initial conditions $x(t_0) \in \mathbb{R}^n$ and all inputs u, there exists a finite time $t \geq t_0$ such that $x(t_0)$ can be uniquely determined from the input u and output y over the interval $[t_0, t]$.

(ii) G is *completely observable* if it is observable from all t_0.

□

6.1 Time-Varying Continuous-Time Relations

Consider the time-varying, continuous-time state-space model

$$\begin{aligned} \dot{x}(t) &= A(t)x(t) + B(t)u(t) \\ y(t) &= C(t)x(t) + D(t)u(t) \end{aligned} \tag{6.1}$$

Definition 6.2 An *observability gramian* of the time-varying, continuous-time model (6.1) is of the form

$$W_o(t, t_0) \triangleq \int\limits_{t_0}^{t} \Phi^T(\tau, t_0) C^T(\tau) C(\tau) \Phi(\tau, t_0) d\tau \tag{6.2}$$

where the state-transition matrix Φ is a solution of (2.4), that is,

$$\frac{d}{dt}\Phi(t, t_0) = A(t)\Phi(t, t_0), \qquad \Phi(t_0, t_0) = I \tag{6.3}$$

□

Remark 6.1 If $A(t)$ is continuous, then there is a unique solution Φ of (6.3), in which case, (6.2) is *the* observability gramian. In practice, it is almost always assumed that the gramian is unique. □

Fact 6.1 Consider the time-varying, continuous-time model (6.1). If $A(t)$ is continuous, then the following statements are equivalent:

(i) For all t_0, there exists a finite $t \geq t_0$ for which the observability gramian (6.2) has full rank.

(ii) The continuous-time model (6.1) is completely observable.

Specifically, the initial condition $x(t_0)$ is given by

$$z(\tau) \triangleq y(\tau) - D(\tau)u(\tau) - C(\tau) \int_{t_0}^{\tau} \Phi(\tau, s)B(s)u(s)ds \tag{6.4}$$

$$x(t_0) = W_o^{-1}(t, t_0) \int_{t_0}^{t} \Phi^T(\tau, t_0)C^T(\tau)z(\tau)d\tau \tag{6.5}$$

Proof ((i)) ⇒ ((ii)): Since $A(t)$ is continuous, then from Fact 2.1, there exists a unique output y of (6.1). Specifically, from (2.3) and (6.4), we have that

$$\begin{aligned}
z(\tau) &= C(\tau)\Phi(\tau, t_0)x(t_0) \\
\Phi^T(\tau, t_0)C^T(\tau)z(\tau) &= \Phi^T(\tau, t_0)C^T(\tau)C(\tau)\Phi(\tau, t_0)x(t_0) \\
\int_{t_0}^{t} \Phi^T(\tau, t_0)C^T(\tau)z(\tau)d\tau &= \int_{t_0}^{t} \Phi^T(\tau, t_0)C^T(\tau)C(\tau)\Phi(\tau, t_0)x(t_0)d\tau \\
\int_{t_0}^{t} \Phi^T(\tau, t_0)C^T(\tau)z(\tau)d\tau &= W_o(t, t_0)x(t_0)
\end{aligned}$$

Therefore, letting t denote the finite time at which $W_o(t, t_0)$ has full rank, we find that the initial condition is given by (6.5).

((ii)) ⇒ ((i)): We use proof by contradiction. Specifically, suppose that (6.1) is observable, and that there exists a t_0 for which the observability gramian (6.2) is singular for all finite $t \geq t_0$. Then for all finite $s \geq t_0$, there exists a nonzero vector $v(s)$ in the nullspace of $W_o(s, t_0)$, that is, for all finite $s \geq t_0$, $W_o(s, t_0)v(s) = 0$. Hence

$$\begin{aligned} v^T(s)W_o(s,t_0)v(s) &= v^T(s) \int_{t_0}^{s} \Phi^T(\tau,t_0)C^T(\tau)C(\tau)\Phi(\tau,t_0)d\tau v(s) \\ &= \int_{t_0}^{s} v^T(s)\Phi^T(\tau,t_0)C^T(\tau)C(\tau)\Phi(\tau,t_0)v(s)d\tau \\ &= \int_{t_0}^{s} \Big\|C(\tau)\Phi(\tau,t_0)v(s)\Big\|_2^2 d\tau = 0 \end{aligned}$$

where, since the norm of a vector is nonnegative, we have that

$$C(\tau)\Phi(\tau,t_0)v(s) = 0 \quad \text{for all finite } s \geq t_0 \text{ and all } \tau \in [t_0, s]$$

that is, there exists a constant nonzero vector $v_\infty = \lim_{s\to\infty} v(s)$ such that

$$C(\tau)\Phi(\tau,t_0)v_\infty = 0 \quad \text{for all finite } \tau \geq t_0 \tag{6.6}$$

Next, let $x(t_0) \neq 0$. Then the output y due to the initial conditions $x(t_0)$ and $x(t_0) + v_\infty$ is given by

$$\begin{aligned} y(t) &= C(t)\Phi(t,t_0)x(t_0) + C(t) \int_{t_0}^{t} \Phi(t,\tau)B(\tau)u(\tau)d\tau + D(t)u(t) \\ &= C(t)\Phi(t,t_0)\Big(x(t_0) + v_\infty\Big) + C(t) \int_{t_0}^{t} \Phi(t,\tau)B(\tau)u(\tau)d\tau + D(t)u(t) \end{aligned}$$

and therefore the addition of v_∞ to the initial condition has no effect on the output y. Hence $x(t_0)$ and $x(t_0) + v_\infty$ cannot be uniquely determined from the input u and output y, which contradicts the fact that the model is observable. □

Fact 6.2 Consider the time-varying, continuous-time, state-space model (6.1). If $A(t)$ is continuous, then the following statements are equivalent:

(i) There exists a finite $t \geq t_0$ for which the observability gramian (6.2) has full rank.

(ii) The continuous-time model (6.1) is observable from t_0.

Proof This proof is a special case of the proof of Fact 6.1. □

6.2 Time-Varying Discrete-Time Relations

Consider the time-varying, discrete-time state-space model

$$\begin{aligned} x(k+1) &= A(k)x(k) + B(k)u(k) \\ y(k) &= C(k)x(k) + D(k)u(k) \end{aligned} \tag{6.7}$$

Definition 6.3 The *observability gramian* of the time-varying, discrete-time state-space model (6.7) is given by

$$W_o(k, k_0) \triangleq \sum_{\kappa=k_0}^{k} \Phi^T(\kappa, k_0)C^T(\kappa)C(\kappa)\Phi(\kappa, k_0), \qquad k \geq k_0 \tag{6.8}$$

where the state-transition matrix Φ is given by (2.18), that is,

$$\Phi(k+1, k_0) = A(k)\Phi(k, k_0) = A(k)\cdots A(k_0), \qquad \Phi(k_0, k_0) = I \tag{6.9}$$

□

Fact 6.3 Consider the time-varying, discrete-time state-space model (6.7). The following statements are equivalent:

(i) For all k_0, there exists a finite $k \geq k_0$ for which the observability gramian (6.8) has full rank.

(ii) The discrete-time model (6.7) is completely observable.

Specifically, the initial condition $x(k_0)$ is given by

$$z(\kappa) \triangleq y(\kappa) - D(\kappa)u(\kappa) - C(\kappa)\sum_{i=k_0}^{\kappa-1} \Phi(\kappa, i+1)B(i)u(i) \tag{6.10}$$

$$x(k_0) = W_o^{-1}(k, k_0) \sum_{\kappa=k_0}^{k} \Phi^T(\kappa, k_0)C^T(\kappa)z(\kappa) \tag{6.11}$$

Proof ((i)) ⇒ ((ii)): From (2.17) and (6.10) we have that

$$\begin{aligned} z(\kappa) &= C(\kappa)\Phi(\kappa, k_0)x(k_0) \\ \Phi^T(\kappa, k_0)C^T(\kappa)z(\kappa) &= \Phi^T(\kappa, k_0)C^T(\kappa)C(\kappa)\Phi(\kappa, k_0)x(k_0) \\ \sum_{\kappa=k_0}^{k} \Phi^T(\kappa, k_0)C^T(\kappa)z(\kappa) &= \sum_{\kappa=k_0}^{k} \Phi^T(\kappa, k_0)C^T(\kappa)C(\kappa)\Phi(\kappa, k_0)x(k_0) \\ \sum_{\kappa=k_0}^{k} \Phi^T(\kappa, k_0)C^T(\kappa)z(\kappa) &= W_o(k, k_0)x(k_0) \end{aligned}$$

Therefore, letting k denote the finite time at which $W_o(k, k_0)$ has full rank, we find that the initial condition is given by (6.11).

((ii)) ⇒ ((i)): We use proof by contradiction. Specifically, suppose that (6.7) is observable, and that there exists a k_0 for which the observability gramian (6.8) is singular for all finite $k \geq k_0$. Then for all finite $s \geq k_0$, there exists a nonzero vector $v(s)$ in the nullspace of $W_o(s, k_0)$, that is, for all finite $s \geq k_0$, $W_o(s, k_0)v(s) = 0$. Hence

$$\begin{aligned} v^T(s)W_o(s,k_0)v(s) &= v^T(s)\sum_{\kappa=k_0}^{s}\Phi^T(\kappa,k_0)C^T(\kappa)C(\kappa)\Phi(\kappa,k_0)v(s) \\ &= \sum_{\kappa=k_0}^{s} v^T(s)\Phi^T(\kappa,k_0)C^T(\kappa)C(\kappa)\Phi(\kappa,k_0)v(s) \\ &= \sum_{\kappa=k_0}^{s} \left\|C(\kappa)\Phi(\kappa,k_0)v(s)\right\|_2^2 = 0 \end{aligned}$$

where, since the norm of a vector is nonnegative, we have that

$$C(\kappa)\Phi(\kappa,k_0)v(s) = 0 \quad \text{for all finite } s \geq k_0 \text{ and all } \kappa \in [k_0, s]$$

that is, there exists a constant nonzero vector $v_\infty = \lim_{s\to\infty} v(s)$ such that

$$C(\kappa)\Phi(\kappa,k_0)v_\infty = 0 \quad \text{for all finite } \kappa \geq k_0$$

Next, let $x(k_0) \neq 0$. Then the output y is identical for both the initial condition $x(k_0)$ and the initial condition $x(k_0) + v_\infty$, that is,

$$\begin{aligned} y(\kappa) &= C(\kappa)\left[\Phi(\kappa,k_0)x(k_0) + \sum_{i=k_0}^{\kappa-1}\Phi(\kappa,i+1)B(i)u(i)\right] + D(\kappa)u(\kappa) \\ &= C(\kappa)\left[\Phi(\kappa,k_0)\left[x(k_0)+v_\infty\right] + \sum_{i=k_0}^{\kappa-1}\Phi(\kappa,i+1)B(i)u(i)\right] + D(\kappa)u(\kappa) \end{aligned}$$

and therefore the addition of v_∞ to the initial condition has no effect on the output y. Hence $x(k_0)$ and $x(k_0)+v_\infty$ cannot be uniquely determined from the input u and output y, which contradicts the fact that the model is observable. □

Fact 6.4 Consider the time-varying, discrete-time state-space model (6.7). The following statements are equivalent:

(i) There exists a finite $k \geq k_0$ for which the observability gramian (6.8) has full rank.

(ii) The discrete-time model (6.7) is observable from k_0.

Proof This proof is a special case of the proof of Fact 6.3. □

6.3 Time-Invariant Relations

In the time-invariant, continuous-time case, the state-space model simplifies to

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) + Du(t) \end{aligned} \tag{6.12}$$

and the observability gramian simplifies to

$$W_o(t, t_0) \triangleq \int_{t_0}^{t} \left(e^{A[\tau - t_0]}\right)^T C^T C e^{A[\tau - t_0]} d\tau \tag{6.13}$$

Similarly, for $k \geq k_0$, the time-invariant, discrete-time model becomes

$$\begin{aligned} x(k+1) &= Ax(k) + Bu(k) \\ y(k) &= Cx(k) + Du(k) \end{aligned} \tag{6.14}$$

and the observability gramian simplifies to

$$W_o(k, k_0) \triangleq \sum_{\kappa=0}^{k-k_0} \left(A^\kappa\right)^T C^T C A^\kappa \tag{6.15}$$

Although the time-invariant gramians are simpler than their time-varying counterparts, they are still much more complicated than they need to be for the simple task of checking observability. Instead, we can simply check the rank of the observability matrix:

Definition 6.4 The *observability matrix* of a time-invariant, continuous or discrete-time state-space model, such as (6.12) or (6.14), is given by

$$\mathcal{O} \triangleq \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} \tag{6.16}$$

where n is the dimension of the matrix A, that is, $A \in \mathbb{R}^{n \times n}$, $C \in \mathbb{R}^{p \times n}$, and $\mathcal{O} \in \mathbb{R}^{pn \times n}$. □

Fact 6.5 The following statements are equivalent:

(i) The time-invariant continuous-time model (6.12) is completely observable.

(ii) The observability matrix $\mathcal{O}$ given in (6.16) has full column rank.

Proof ((i)) $\Rightarrow$ ((ii)): From Fact 6.1, if (6.12) is completely observable, then for all t_0, there exists a finite $t \geq t_0$ for which the observability gramian has full rank. Furthermore, from Corollary 5.1 we have that

$$e^{A(\tau - t_0)} = p_0(\tau - t_0)I + p_1(\tau - t_0)A + \cdots + p_{n-1}(\tau - t_0)A^{n-1}$$

Thus letting

$$\mathcal{P}(\tau - t_0) \triangleq \begin{bmatrix} p_0(\tau - t_0)I & \cdots & p_{n-1}(\tau - t_0)I \end{bmatrix}$$

it follows that $Ce^{A(\tau - t_0)} = \mathcal{P}(\tau - t_0)\mathcal{O}$, and therefore

$$W_o(t, t_0) = \mathcal{O}^T \left(\int_{t_0}^{t} \mathcal{P}^T(\tau - t_0)\mathcal{P}(\tau - t_0) d\tau \right) \mathcal{O}$$

Finally, since there exists a finite $t \geq t_0$ for which the observability gramian has full rank, the observability matrix must have full column rank.

((ii)) $\Rightarrow$ ((i)): Suppose that the observability matrix has full rank, but the system is not completely controllable. Then from Fact 6.1, there exists a t_0 for which the observability gramian is singular for all finite $t \geq t_0$. Specifically, referring to (6.6) in the proof of Fact 6.1, for all finite times $\tau \geq t_0$, there exists a nonzero v_∞ satisfying

$$Ce^{A[\tau - t_0]}v_\infty = 0 \tag{6.17}$$

Hence differentiating (6.17) i times with respect to τ, we find that

$$CA^i e^{A[\tau - t_0]}v_\infty = 0$$

and therefore

$$\mathcal{O}e^{A[\tau - t_0]}v_\infty = 0$$

However, since the matrix exponential has full rank (Fact 5.9) and $\mathcal{O}$ has full column rank, it follows that $v_\infty = 0$, which contradicts the fact that there exists a nonzero v_∞ satisfying (6.17). □

Fact 6.6 The following statements are equivalent:

(i) The time-invariant discrete-time model (6.14) is completely observable.

(ii) The observability matrix $\mathcal{O}$ given in (6.16) has full column rank.

Specifically, the initial condition $x(k_0)$ is given by

$$z(k_0+\kappa) \triangleq y(k_0+\kappa) - Du(k_0+\kappa) - C\sum_{i=0}^{\kappa-1} A^{\kappa-1-i}Bu(i) \tag{6.18}$$

$$x(k_0) = \left(\mathcal{O}^T\mathcal{O}\right)^{-1}\mathcal{O}^T \begin{bmatrix} z(k_0) \\ \vdots \\ z(k_0+n-1) \end{bmatrix} \tag{6.19}$$

Proof From Corollary 5.1, recall that for all $\kappa \geq n$, there exists $\tilde{p}_{\kappa,0}, \ldots, \tilde{p}_{\kappa,n-1}$ such that

$$CA^\kappa = \begin{bmatrix} \tilde{p}_{\kappa,0}I & \cdots & \tilde{p}_{\kappa,n-1}I \end{bmatrix}\mathcal{O} = \tilde{\mathcal{P}}_\kappa\mathcal{O}$$

Therefore, if $k-k_0 \geq n-1$, then the discrete-time observability gramian is of the form

$$\begin{aligned} W_o(k,k_0) &= \sum_{\kappa=0}^{n-1}\left(A^\kappa\right)^T C^TCA^\kappa + \sum_{\kappa=n}^{k-k_0}\left(A^\kappa\right)^T C^TCA^\kappa \\ &= \mathcal{O}^T\mathcal{O} + \sum_{\kappa=n}^{k-k_0}\mathcal{O}^T\tilde{\mathcal{P}}_\kappa^T\tilde{\mathcal{P}}_\kappa\mathcal{O} = \mathcal{O}^T\left(I + \sum_{\kappa=n}^{k-k_0}\tilde{\mathcal{P}}_\kappa^T\tilde{\mathcal{P}}_\kappa\right)\mathcal{O} \end{aligned}$$

Furthermore, since $\tilde{\mathcal{P}}_\kappa^T\tilde{\mathcal{P}}_\kappa$ is positive semi-definite, $I + \sum_{\kappa=n}^{k-k_0}\tilde{\mathcal{P}}_\kappa^T\tilde{\mathcal{P}}_\kappa$ is positive definite, that is, it has full rank. Hence for all k_0 and all $k \geq k_0+n-1$, $W_o(k,k_0)$ has full rank if and only if $\mathcal{O}$ has full column rank. Therefore, from Fact 6.3, we have that ((i)) and ((ii)) are equivalent.

Finally, if the observability matrix has full column rank, then letting $k = k_0+n-1$ in (6.11), we find (6.19). □

SECTION 7

Linearity

Using the definition of a state (Definition 4.1), we introduce the following notion of a *linear system*:

Definition 7.1 For $i = 1, 2$, let y_i denote the output of the system G due to the input u_i and initial condition $x_i(t_0)$. Then G is *linear* if both of the following conditions hold:

1) (**Scalability**) For all $\alpha \in \mathbb{R}$, all initial conditions $x_1(t_0)$, and all inputs u_1, the output of G due to the input αu_1 and initial condition $\alpha x_1(t_0)$ is αy_1.

2) (**Additivity**) For all initial conditions $x_1(t_0)$ and $x_2(t_0)$, and for all inputs u_1 and u_2, the output of G due to the input $u_1 + u_2$ and initial condition $x_1(t_0) + x_2(t_0)$ is $y_1 + y_2$.

Next, we give conditions for which the state-space models are linear, where we are particularly concerned with the four forms:

Time-varying Continuous-time	$\dot{x}(t) = A(t)x(t) + B(t)u(t)$ $y(t) = C(t)x(t) + D(t)u(t)$
Time-invariant Continuous-time	$\dot{x}(t) = Ax(t) + Bu(t)$ $y(t) = Cx(t) + Du(t)$
Time-varying Discrete-time	$x(k+1) = A(k)x(k) + B(k)u(k)$ $y(k) = C(k)x(k) + D(k)u(k)$
Time-invariant Discrete-time	$x(k+1) = Ax(k) + Bu(k)$ $y(k) = Cx(k) + Du(k)$

Table 1: Common state-space models.

Theorem 7.1 Consider the state-space models shown in Table 1. The time-invariant models, along with the time-varying discrete-time model, are linear. If $A(t)$ is continuous, then the time-varying, continuous-time model is also linear.

Proof The time-invariant models are special cases of the time-varying models. Hence we only need to show that the time-varying models are linear.

We start with the time-varying, continuous-time model. Since $A(t)$ is continuous, then from Fact 2.1, there exists a unique state x. Hence from (2.3), we have that the output y due to the input αu and initial condition $\alpha x(t_0)$ is given by

$$C(t)\Phi(t,t_0)\Big[\alpha x(t_0)\Big] + C(t)\int\limits_{t_0}^{t} \Phi(t,\tau)B(\tau)\Big[\alpha u(\tau)\Big]d\tau + D(t)\Big[\alpha u(t)\Big]$$

$$= \alpha \left[C(t)\Phi(t,t_0)x(t_0) + C(t)\int\limits_{t_0}^{t} \Phi(t,\tau)B(\tau)u(\tau)d\tau + D(t)u(t)\right] = \alpha y(t)$$

Therefore the time-varying, continuous-time model satisfies the scalability property of Definition 7.1. Furthermore, letting y_i denote the output due to the input u_i and initial condition $x_i(t_0)$, we have that the output due to the input $u_1 + u_2$ and initial condition $x_1(t_0) + x_2(t_0)$ is given by

$$\begin{aligned}
&C(t)\Phi(t,t_0)\Big[x_1(t_0) + x_2(t_0)\Big] + D(t)\Big[u_1(t) + u_2(t)\Big] \\
&+ C(t)\int\limits_{t_0}^{t} \Phi(t,\tau)B(\tau)\Big[u_1(\tau) + u_2(\tau)\Big]d\tau \\
&= \left[C(t)\Phi(t,t_0)x_1(t_0) + C(t)\int\limits_{t_0}^{t} \Phi(t,\tau)B(\tau)u_1(\tau)d\tau + D(t)u_1(t)\right] \\
&+ \left[C(t)\Phi(t,t_0)x_2(t_0) + C(t)\int\limits_{t_0}^{t} \Phi(t,\tau)B(\tau)u_2(\tau)d\tau + D(t)u_2(t)\right] \\
&= y_1(t) + y_2(t)
\end{aligned}$$

and therefore the time-varying, continuous-time model also satisfies the additivity property of Definition 7.1. Hence the time-varying, continuous-time model is linear.

Finally, following the same procedure as in the continuous-time case, we find that the time-varying, discrete-time state-space model also satisfies the scalability and additivity properties of Definition 7.1. Therefore it is also linear. □

SECTION 8

Operators

By now you will have probably noticed that the continuous-time and discrete-time models we presented looked very similar. For instance, the time-varying continuous and discrete-time state-space models (2.1) and (2.15) are both of the form

$$\begin{aligned} \boldsymbol{\rho}\big[x(t)\big] &= A(t)x(t) + B(t)u(t) \\ y(t) &= C(t)x(t) + D(t)u(t) \end{aligned} \tag{8.1}$$

where the continuous-time models use the differentiation operator $\boldsymbol{\rho} \triangleq d/dt$, and the discrete-time models use the forward-shift operator $\boldsymbol{\rho} \triangleq \mathbf{q}$, that is,

$$\mathbf{q}\big[x(k)\big] = x(k+1) \tag{8.2}$$

$$\mathbf{q}^n\big[x(k)\big] = x(k+n) \tag{8.3}$$

However, although the models appear structurally similar when expressed in the generic operator $\boldsymbol{\rho}$, they have very different solutions. For instance, the solution of (8.1) when $\boldsymbol{\rho} \triangleq d/dt$ is given by (2.2)-(2.3), whereas the solution of (8.1) when $\boldsymbol{\rho} \triangleq \mathbf{q}$ is given by (2.16)-(2.17).

Nonetheless, as we will see in coming chapters, operator notation is very convenient for expressing higher-order differential equations and time-series. For instance, consider the differential equation

$$\begin{aligned} &y^{(n)}(t) + \alpha_{n-1}(t)y^{(n-1)}(t) + \cdots + \alpha_0(t)y(t) \\ &= \beta_n(t)u^{(n)}(t) + \beta_{n-1}(t)u^{(n-1)}(t) + \cdots + \beta_0(t)u(t) \end{aligned} \tag{8.4}$$

Then using operator notation, we can write (8.4) compactly as

$$A\Big(t, \frac{d}{dt}\Big)y(t) = B\Big(t, \frac{d}{dt}\Big)u(t) \tag{8.5}$$

where

$$\begin{aligned} A\Big(t, \frac{d}{dt}\Big) &\triangleq \quad\;\; \frac{d}{dt}^{n} + \alpha_{n-1}(t)\frac{d}{dt}^{n-1} + \cdots + \alpha_1(t)\frac{d}{dt} + \alpha_0(t) \\ B\Big(t, \frac{d}{dt}\Big) &\triangleq \beta_n(t)\frac{d}{dt}^{n} + \beta_{n-1}(t)\frac{d}{dt}^{n-1} + \cdots + \beta_1(t)\frac{d}{dt} + \beta_0(t) \end{aligned} \tag{8.6}$$

8.1 Operators are not Real or Complex Numbers

In the next few sections, we illustrate the sometimes counter-intuitive "gotchas" that may lead to confusion when dealing with polynomials of operators, such as (8.6). The most important point is that operators are not numbers. Instead, operators act on signals, where the result of the operation is a new signal. For instance, in the equation

$$\frac{d}{dt}\left[t^3\right] = 3t^2 \tag{8.7}$$

the operator d/dt operates on the signal t^2, yielding the new signal $3t^2$. Similarly, (8.5) should be read as:

(i) A operates on the signal y, yielding the new signal $y_A \triangleq Ay$.

(ii) B operates on the signal u, yielding the new signal $u_B \triangleq Bu$.

(iii) The signals y_A and u_B are equal.

Remark 8.1 The operator polynomials A and B given in (8.6) can also be regarded as simply *operators*, that is, the distinction between operator polynomials and operators is small, and only useful if we can somehow exploit the polynomial structure. □

8.1.1 Multiplication of Operator Polynomials

Since the result of an operator acting on a signal is a new signal, we can also apply an operator to the new signal. For instance, let

$$C(t, \boldsymbol{\rho}) \triangleq \gamma_n(t)\boldsymbol{\rho}^n + \gamma_{n-1}(t)\boldsymbol{\rho}^{n-1} + \cdots + \gamma_1(t)\boldsymbol{\rho} + \gamma_0(t) \tag{8.8}$$

Then from (8.5), we have that

$$C(t, \boldsymbol{\rho})A(t, \boldsymbol{\rho})y(t) = C(t, \boldsymbol{\rho})B(t, \boldsymbol{\rho})u(t) \tag{8.9}$$

where the left and right-hand sides are evaluated by computing the intermediate signals $y_A(t) \triangleq A(t, \boldsymbol{\rho})y(t)$ and $u_B(t) \triangleq B(t, \boldsymbol{\rho})u(t)$, and then applying $C(t, \boldsymbol{\rho})$ to the signals y_A and u_B. We formally define this operation with the following rule:

Definition 8.1 Let the signals y_A and z be given by

$$y_A(t) \triangleq A(t, \boldsymbol{\rho})y(t), \qquad z(t) \triangleq C(t, \boldsymbol{\rho})y_A(t) \tag{8.10}$$

Then we write

$$z(t) = C(t, \boldsymbol{\rho}) A(t, \boldsymbol{\rho}) y(t) \tag{8.11}$$

□

The rule given in Definition 8.1 has important consequences for how we multiply operator polynomials, as we demonstrate in the following examples:

Example 8.1 Let

$$C\left(t, \frac{d}{dt}\right) \triangleq \frac{d}{dt} - \alpha(t), \qquad A\left(t, \frac{d}{dt}\right) \triangleq \frac{d}{dt} + \alpha(t) \tag{8.12}$$

Then it follows that

$$\begin{aligned} C\left(t, \frac{d}{dt}\right) A\left(t, \frac{d}{dt}\right) y(t) &= C\left(t, \frac{d}{dt}\right) \Big(\dot{y}(t) + \alpha(t) y(t)\Big) \\ &= \Big(\ddot{y}(t) + \dot{\alpha}(t) y(t) + \alpha(t)\dot{y}(t)\Big) \\ &\qquad - \Big(\alpha(t)\dot{y}(t) + \alpha^2(t) y(t)\Big) \\ &= \ddot{y}(t) + \dot{\alpha}(t) y(t) - \alpha^2(t) y(t) \end{aligned}$$

□

Example 8.2 Let

$$C\big(k, \mathbf{q}\big) \triangleq \mathbf{q} - \alpha(k), \qquad A\big(k, \mathbf{q}\big) \triangleq \mathbf{q} + \alpha(k) \tag{8.13}$$

Then it follows that

$$\begin{aligned} C\big(k, \mathbf{q}\big) A\big(k, \mathbf{q}\big) y(k) &= C\big(k, \mathbf{q}\big)\big(y(k+1) + \alpha(k) y(k)\big) \\ &= \Big(y(k+2) + \alpha(k+1) y(k+1)\Big) \\ &\qquad - \Big(\alpha(k) y(k+1) + \alpha^2(k) y(k)\Big) \\ &= y(k+2) + \Big(\alpha(k+1) - \alpha(k)\Big) y(k+1) - \alpha^2(k) y(k) \end{aligned} \tag{8.14}$$

□

From Examples 8.1 and 8.2 we see that operator polynomials behave differently than polynomials in real or complex numbers. Specifically, note that if A and C in (8.12) and (8.13) were polynomials in the complex number $s \in \mathbb{C}$, then

$$C\big(t, s\big)A\big(t, s\big)y(t) = s^2 y(t) - \alpha^2(t)y(t) \tag{8.15}$$

which is quite different than what we found in the examples.

The important result demonstrated in Examples 8.1 and 8.2 is that operators do not commute with signals, as you already know from your experience with differentiation. Specifically, recall that

$$\alpha(t)\frac{d}{dt}\Big[y(t)\Big] \neq \frac{d}{dt}\Big[\alpha(t)y(t)\Big] \tag{8.16}$$

Hence the product of $C(t, d/dt)$ and $A(t, d/dt)$ in (8.12) should be written as

$$\left(\frac{d}{dt} - \alpha(t)\right)\left(\frac{d}{dt} + \alpha(t)\right) = \frac{d^2}{dt^2} - \alpha(t)\frac{d}{dt} + \frac{d}{dt}\alpha(t) - \alpha^2(t) \tag{8.17}$$

where the term $\frac{d}{dt}\alpha(t)$ does not mean $\dot{\alpha}(t)$. Instead, it should instead be interpreted as an operator, whose meaning is clarified with the following definition:

Definition 8.2 Let y and $\alpha_1, \alpha_2, \ldots, \alpha_n$ denote signals, and let $\boldsymbol{\rho}$ denote an operator such as the differentiation operator d/dt. Also, let $i_1, \ldots, i_n$ be nonnegative integers, and suppose that the operator polynomial $A(t, \boldsymbol{\rho})$ satisfes

$$A(t, \boldsymbol{\rho})y(t) = \alpha_n(t)\boldsymbol{\rho}^{i_n}\left[\alpha_{n-1}(t)\boldsymbol{\rho}^{i_{n-1}}\left[\cdots \alpha_1(t)\boldsymbol{\rho}^{i_1}\Big[y(t)\Big]\right]\right] \tag{8.18}$$

that is,

$$A(t, \boldsymbol{\rho})y(t) = y_n(t) \tag{8.19}$$

where $y_0(t) \triangleq y(t)$ and for $i = 1, \ldots, n$

$$y_i(t) \triangleq \alpha_i(t)\left(\boldsymbol{\rho}^{i_i}\Big[y_{i-1}(t)\Big]\right) \tag{8.20}$$

Then we write

$$A(t, \boldsymbol{\rho}) = \alpha_n(t)\boldsymbol{\rho}^{i_n}\alpha_{n-1}(t)\boldsymbol{\rho}^{i_{n-1}} \cdots \alpha_1(t)\boldsymbol{\rho}^{i_1} \tag{8.21}$$

□

Hence context is very important. When multiplying operator polynomials, it is often convenient to use the abstract operator $\boldsymbol{\rho}$ which does not commute. For instance, (8.17) would be written as

$$\Big(\boldsymbol{\rho} - \alpha(t)\Big)\Big(\boldsymbol{\rho} + \alpha(t)\Big) = \boldsymbol{\rho}^2 - \alpha(t)\boldsymbol{\rho} + \boldsymbol{\rho}\alpha(t) - \alpha^2(t) \tag{8.22}$$

Then, when we apply this expression to a signal, such as y, we have that

$$\Big(\boldsymbol{\rho} - \alpha(t)\Big)\Big(\boldsymbol{\rho} + \alpha(t)\Big)y(t) = \boldsymbol{\rho}^2 y(t) - \alpha(t)\boldsymbol{\rho}y(t) + \boldsymbol{\rho}\alpha(t)y(t) - \alpha^2(t)y(t)$$

where, if $\boldsymbol{\rho}$ denotes the differentiation operator, we find that

$$\begin{aligned}&\Big(\boldsymbol{\rho} - \alpha(t)\Big)\Big(\boldsymbol{\rho} + \alpha(t)\Big)y(t)\\ &\qquad = \ddot{y}(t) - \alpha(t)\dot{y}(t) + \Big(\dot{\alpha}(t)y(t) + \alpha(t)\dot{y}(t)\Big) - \alpha^2(t)y(t)\end{aligned}$$

which agrees with Example 8.1.

8.1.2 Comparisons

Comparisons between operators or functions of operators are only valid when making statements about equality or inequality. For instance,

$$\begin{aligned}&\text{(Valid Expression)}: \qquad && 2\frac{d}{dt} \neq \frac{d}{dt}\\ &\text{(Valid Expression)}: \qquad && 2\frac{d}{dt} = 2\frac{d}{dt}\end{aligned} \tag{8.23}$$

where we formally define equality to mean:

Definition 8.3 Let $\boldsymbol{\rho}$ be an operator, such as the differentiation operator d/dt, and let $f(t, \boldsymbol{\rho})$ and $g(t, \boldsymbol{\rho})$ be operator polynomials in $\boldsymbol{\rho}$. Also, let $\mathcal{S}$ denote the set of all signals that $\boldsymbol{\rho}$ can operate on. Then we say that $f(t, \boldsymbol{\rho}) = g(t, \boldsymbol{\rho})$ if

$$f(t, \boldsymbol{\rho})s(t) = g(t, \boldsymbol{\rho})s(t) \qquad \text{for all } t \text{ and all } s \in \mathcal{S} \tag{8.24}$$

□

From this definition, it should be clear that *greater than* and *less than* comparisons of operators are not valid. For instance,

$$\text{(Invalid Expression)}: \qquad 2\frac{d}{dt} \geq \frac{d}{dt} \tag{8.25}$$

is invalid since the equation can yield different results depending on the signal it is applied to. For instance, if the operators in (8.25) act on the signal t^3, we find that

$$2\frac{d}{dt}\Big[t^3\Big] = 6t^2 \geq 3t^2 = \frac{d}{dt}\Big[t^3\Big] \tag{8.26}$$

whereas, if they operate on the signal $-t^3$, we find that

$$2\frac{d}{dt}\Big[-t^3\Big] = -6t^2 \leq -3t^2 = \frac{d}{dt}\Big[-t^3\Big] \tag{8.27}$$

Another consequence of Definition 8.3 is that, in general,

$$C(t,\boldsymbol{\rho})A(t,\boldsymbol{\rho}) \neq A(t,\boldsymbol{\rho})C(t,\boldsymbol{\rho}) \tag{8.28}$$

even in the case where A and C are scalar polynomials.

8.1.3 Division and Cancellation

Division by an operator is invalid. For instance,

$$\text{(Invalid Expression)}: \qquad \frac{1}{d/dt}\Big[y(t)\Big] = 5 \tag{8.29}$$

is unclear. Do we mean integration? or perhaps we mean $y(t) = \frac{d}{dt}[5]$. We will never divide by an operator in this book. If you intend to use such operations, then define what division means, and make sure that your notation is consistent.

Furthermore, cancellation of operators is invalid, that is,

$$A\left(t, \frac{d}{dt}\right)y(t) = A\left(t, \frac{d}{dt}\right)u(t) \quad \not\Rightarrow \quad y(t) = u(t) \tag{8.30}$$

For instance, if $A\big(t, d/dt\big) \triangleq d/dt$, then $y(t) = u(t) + 5$ is a solution of $A\big(t, d/dt\big)y(t) = A\big(t, d/dt\big)u(t)$ where $y(t) \neq u(t)$.

8.1.4 Time-Invariant Operator Polynomials

Unlike time-varying operator polynomials, time-invariant polynomials in the differentiation d/dt and forward-shift $\mathbf{q}$ operators behave like normal polynomials under multiplication. This is because constants commute with these operators. Specifically, we mean polynomials of the form

$$B(\boldsymbol{\rho}) = \beta_n \boldsymbol{\rho}^n + \cdots + \beta_1 \boldsymbol{\rho} + \beta_0 \tag{8.31}$$

where $\beta_0, \ldots, \beta_n \in \mathbb{R}^{n \times m}$ are constant matrices. The multiplication of time-invariant operator polynomials is considered in the following examples:

Example 8.3 Let α be constant and let

$$C\left(\frac{d}{dt}\right) \triangleq \frac{d}{dt} - \alpha, \qquad A\left(\frac{d}{dt}\right) \triangleq \frac{d}{dt} + \alpha \tag{8.32}$$

Then it follows that

$$\begin{aligned} C\left(\frac{d}{dt}\right) A\left(\frac{d}{dt}\right) y(t) &= C\left(\frac{d}{dt}\right)\Big(\dot{y}(t) + \alpha y(t)\Big) \\ &= \Big(\ddot{y}(t) + \alpha \dot{y}(t)\Big) - \Big(\alpha \dot{y}(t) + \alpha^2 y(t)\Big) \\ &= \ddot{y}(t) - \alpha^2 y(t) \end{aligned} \tag{8.33}$$

□

Example 8.4 Let α be constant and let

$$C(\mathbf{q}) \triangleq \mathbf{q} - \alpha, \qquad A(\mathbf{q}) \triangleq \mathbf{q} + \alpha \tag{8.34}$$

Then it follows that

$$\begin{aligned} C(\mathbf{q}) A(\mathbf{q}) y(k) &= C(\mathbf{q}) \big(y(k+1) + \alpha y(k)\big) \\ &= \Big(y(k+2) + \alpha y(k+1)\Big) - \Big(\alpha y(k+1) + \alpha^2 y(k)\Big) \\ &= y(k+2) - \alpha^2 y(k) \end{aligned} \tag{8.35}$$

□

Of course, one still needs to be careful in the matrix about whether or not certain matrices commute:

Example 8.5 Let $A_0, A_1 \in \mathbb{R}^{p \times p}$ and let

$$C\left(\frac{d}{dt}\right) \triangleq A_1 \frac{d}{dt} - A_0, \qquad A\left(\frac{d}{dt}\right) \triangleq A_1 \frac{d}{dt} + A_0 \tag{8.36}$$

Then it follows that

$$\begin{aligned} C\left(\frac{d}{dt}\right)A\left(\frac{d}{dt}\right)y(t) &= C\left(\frac{d}{dt}\right)\Big(A_1\dot{y}(t)+A_0y(t)\Big) \qquad (8.37) \\ &= A_1\Big(A_1\ddot{y}(t)+A_0\dot{y}(t)\Big)-A_0\Big(A_1\dot{y}(t)+A_0y(t)\Big) \\ &= A_1^2\ddot{y}(t)+\Big(A_1A_0-A_0A_1\Big)\dot{y}(t)-A_0^2y(t) \end{aligned}$$

and hence $A_1A_0 - A_0A_1 = 0$ if and only if A_1 and A_0 commute. □

SECTION 9

Polynomial Matrix Models

9.1 Time-Varying Continuous-Time Models

Time-varying, continuous-time polynomial matrix models are of the form

$$\begin{aligned} & y^{(n)}(t) + \alpha_{n-1}(t)y^{(n-1)}(t) + \cdots + \alpha_0(t)y(t) \\ &= \beta_n(t)u^{(n)}(t) + \beta_{n-1}(t)u^{(n-1)}(t) + \cdots + \beta_0(t)u(t) \end{aligned} \tag{9.1}$$

where $y \in \mathbb{R}^p$, $u \in \mathbb{R}^m$, $\alpha_0, \ldots, \alpha_{n-1} \in \mathbb{R}^{p \times p}$, and $\beta_0, \ldots, \beta_n \in \mathbb{R}^{p \times m}$.

Remark 9.1 Models of the form (9.1) are called *polynomial matrix* models since the inputs and outputs are allowed to be vectors in $\mathbb{R}^m$ and $\mathbb{R}^p$, respectively. Hence the coefficients $\alpha_0, \ldots, \alpha_{n-1}$, and $\beta_0, \ldots, \beta_n$ are matrices. □

Remark 9.2 Continuous-time polynomial matrix models are sometimes called *sets of ordinary differential equations.* □

In previous chapters, we introduced the concepts of reachability, controllability, observability, and linearity. Furthermore, we determined the conditions a state-space model must satisfy in order to have these properties. Here we will determine the conditions a polynomial matrix model must satisfy in order to also have these properties. Specifically, we will accomplish this by developing state-space models which produce the same output y as the polynomial matrix models they are derived from, after which we will show how to apply the results of the previous chapters. The reason to perform the intermediate step of constructing a state-space model with the same output is two-fold:

(i) In many of the definitions we previously introduced, you will find that the concept of a state (Definition 4.1) and initial condition are central to all of definitions. Hence in a lot of cases, it "makes sense" to first convert to a state-space model.

(ii) State-space models have the benefit that it is "easy" to write the state x and output y directly as a function of time, as in (2.2) and (2.3).

9.1.1 To State-Space

Here we show how to construct a time-varying continuous-time state-space model of the form

$$\begin{aligned} \dot{x}(t) &= A(t)x(t) + B(t)u(t) \\ y(t) &= C(t)x(t) + D(t)u(t) \end{aligned} \tag{9.2}$$

with the same output y as the time-varying continuous-time polynomial matrix model (9.1). Specifically, we decompose the conversion process into two steps:

1) Convert the polynomial matrix model (9.1) into the modified form

$$\begin{aligned} &\frac{d^n}{dt^n}\Big[y(t)\Big] + \frac{d^{n-1}}{dt^{n-1}}\Big[\tilde{\alpha}_{n-1}(t)y(t)\Big] + \cdots + \Big[\tilde{\alpha}_0(t)y(t)\Big] \\ = &\frac{d^n}{dt^n}\Big[\tilde{\beta}_n(t)u(t)\Big] + \frac{d^{n-1}}{dt^{n-1}}\Big[\tilde{\beta}_{n-1}(t)u(t)\Big] + \cdots + \Big[\tilde{\beta}_0(t)u(t)\Big] \end{aligned} \tag{9.3}$$

2) Convert the modified polynomial matrix model (9.3) into the state-space model (9.2).

Converting to the Modified Polynomial Matrix Model

To convert the traditional polynomial matrix model (9.1) into the modified form (9.3), we need to develop the relationships between the coefficients α_i and $\tilde{\alpha}_i$, and β_i and $\tilde{\beta}_i$. To accomplish this, note that the terms in (9.3) are of the form

$$\frac{d^i}{dt^i}\Big[\tilde{\alpha}_i(t)y(t)\Big] = \sum_{k=0}^{i} \begin{pmatrix} i \\ k \end{pmatrix} \tilde{\alpha}_i^{(i-k)}(t)y^{(k)}(t) \tag{9.4}$$

where $\begin{pmatrix} i \\ k \end{pmatrix}$ denotes the binomial coefficient. Hence rewriting (9.3) in summation form, we have that

$$y^{(n)}(t) + \sum_{i=0}^{n-1} \frac{d^i}{dt^i}\Big[\tilde{\alpha}_i(t)y(t)\Big] = \sum_{i=0}^{n} \frac{d^i}{dt^i}\Big[\tilde{\beta}_i(t)u(t)\Big] \tag{9.5}$$

and therefore, using (9.4), we find that

$$y^{(n)}(t) + \sum_{i=0}^{n-1}\sum_{k=0}^{i} \begin{pmatrix} i \\ k \end{pmatrix} \tilde{\alpha}_i^{(i-k)}(t)y^{(k)}(t) = \sum_{i=0}^{n}\sum_{k=0}^{i} \begin{pmatrix} i \\ k \end{pmatrix} \tilde{\beta}_i^{(i-k)}(t)u^{(k)}(t) \tag{9.6}$$

Furthermore, reversing the order of the summations, it follows that

$$y^{(n)}(t)+\sum_{k=0}^{n-1}\sum_{i=k}^{n-1}\binom{i}{k}\tilde{\alpha}_i^{(i-k)}(t)y^{(k)}(t)=\sum_{k=0}^{n}\sum_{i=k}^{n}\binom{i}{k}\tilde{\beta}_i^{(i-k)}(t)u^{(k)}(t) \tag{9.7}$$

and thus comparing (9.1) and (9.7), we have that

$$\alpha_k(t)=\sum_{i=k}^{n-1}\binom{i}{k}\tilde{\alpha}_i^{(i-k)}(t), \quad k=0,\ldots,n-1 \tag{9.8}$$

$$\beta_k(t)=\sum_{i=k}^{n}\binom{i}{k}\tilde{\beta}_i^{(i-k)}(t), \quad k=0,\ldots,n \tag{9.9}$$

Unfortunately, the relationships (9.8) and (9.9) are not immediately useful since we need expressions for $\tilde{\alpha}$ and $\tilde{\beta}$ in terms of α and β, that is we need the inverse relationships of (9.8) and (9.9). To develop the inverse relationships, note that (9.8) and (9.9) can be written as

$$\begin{bmatrix}\beta_0(t)\\ \beta_1(t)\\ \vdots\\ \beta_n(t)\end{bmatrix}=\begin{bmatrix}\binom{0}{0}\frac{d^0}{dt^0} & \binom{1}{0}\frac{d^1}{dt^1} & \cdots & \binom{n}{0}\frac{d^n}{dt^n}\\ 0 & \binom{1}{1}\frac{d^0}{dt^0} & \cdots & \binom{n}{1}\frac{d^{n-1}}{dt^{n-1}}\\ \vdots & \ddots & \ddots & \vdots\\ 0 & \cdots & 0 & \binom{n}{n}\frac{d^0}{dt^0}\end{bmatrix}\begin{bmatrix}\tilde{\beta}_0(t)\\ \tilde{\beta}_1(t)\\ \vdots\\ \tilde{\beta}_n(t)\end{bmatrix} \tag{9.10}$$

Hence computing the inverse, we find that

$$\tilde{\alpha}_k(t)=\sum_{i=k}^{n-1}(-1)^{i-k}\binom{i}{k}\alpha_i^{(i-k)}(t), \quad k=0,\ldots,n-1 \tag{9.11}$$

$$\tilde{\beta}_k(t)=\sum_{i=k}^{n}(-1)^{i-k}\binom{i}{k}\beta_i^{(i-k)}(t), \quad k=0,\ldots,n \tag{9.12}$$

Therefore, the polynomial matrix model (9.1) can be written in the modified form (9.3), where $\tilde{\alpha}_k$ and $\tilde{\beta}_k$ are given by (9.11) and (9.12).

From the Modified Form to State-Space

Next, we show how to construct a state-space model with the same output as the modified polynomial matrix model (9.3). Specifically, let y be the output of the state-space model (9.2), where

$$A(t) \triangleq \begin{bmatrix} 0_{p\times p} & \cdots & \cdots & 0_{p\times p} & -\tilde{\alpha}_0(t) \\ I_p & \ddots & & \vdots & -\tilde{\alpha}_1(t) \\ 0_{p\times p} & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0_{p\times p} & -\tilde{\alpha}_{n-2}(t) \\ 0_{p\times p} & \cdots & 0_{p\times p} & I_p & -\tilde{\alpha}_{n-1}(t) \end{bmatrix} \in \mathbb{R}^{np\times np} \tag{9.13}$$

$$B(t) \triangleq \begin{bmatrix} \tilde{\beta}_0(t) - \tilde{\alpha}_0(t)\tilde{\beta}_n(t) \\ \vdots \\ \tilde{\beta}_{n-1}(t) - \tilde{\alpha}_{n-1}(t)\tilde{\beta}_n(t) \end{bmatrix} \in \mathbb{R}^{np\times m} \tag{9.14}$$

$$C(t) \triangleq \begin{bmatrix} 0_{p\times p} & \cdots & 0_{p\times p} & I_p \end{bmatrix} \in \mathbb{R}^{p\times np} \tag{9.15}$$

$$D(t) \triangleq \tilde{\beta}_n(t) \in \mathbb{R}^{p\times m} \tag{9.16}$$

that is,

$$x(t) \triangleq \begin{bmatrix} x_1(t) \\ \vdots \\ x_n(t) \end{bmatrix} \in \mathbb{R}^{np\times 1} \tag{9.17}$$

$$\dot{x}_1(t) \triangleq \quad -\tilde{\alpha}_0(t)x_n(t) + \Big[\tilde{\beta}_0(t) - \tilde{\alpha}_0(t)\tilde{\beta}_n(t)\Big]u(t) \tag{9.18}$$

$$\dot{x}_{i+1}(t) \triangleq x_i(t) - \tilde{\alpha}_i(t)x_n(t) + \Big[\tilde{\beta}_i(t) - \tilde{\alpha}_i(t)\tilde{\beta}_n(t)\Big]u(t), \quad i = 1, \ldots, n-1$$

$$y(t) \triangleq x_n(t) + \tilde{\beta}_n(t)u(t) \tag{9.19}$$

Then we will show that y is also an output of the modified polynomial matrix model (9.3). Specifically, from (9.18)-(9.19), we have that

$$\dot{x}_1(t) = \tilde{\beta}_0(t)u(t) - \tilde{\alpha}_0(t)y(t) \tag{9.20}$$

$$x_{i-1}(t) = \dot{x}_i(t) + \tilde{\alpha}_{i-1}(t)y(t) - \tilde{\beta}_{i-1}(t)u(t), \quad i = 2, \ldots, n \tag{9.21}$$

Hence from (9.19) and (9.21), it follows that

$$x_{n-1}(t) = \frac{d}{dt}\Big[y(t) - \tilde{\beta}_n(t)u(t)\Big] + \tilde{\alpha}_{n-1}(t)y(t) - \tilde{\beta}_{n-1}(t)u(t) \tag{9.22}$$

Furthermore, recursively substituting into (9.21), we find that $x_i(t)$ is given by

$$x_i(t) = \sum_{k=0}^{n-i} \frac{d^k}{dt^k}\left[\tilde{\alpha}_{i+k}(t)y(t) - \tilde{\beta}_{i+k}(t)u(t)\right], \quad i = 1, \ldots, n \tag{9.23}$$

where $\tilde{\alpha}_n = I_p$. Therefore, differentiating (9.23) and evaluating at $i = 1$ yields

$$\dot{x}_1(t) = \sum_{k=1}^{n} \frac{d^k}{dt^k}\left[\tilde{\alpha}_k(t)y(t) - \tilde{\beta}_k(t)u(t)\right] \tag{9.24}$$

Finally, comparing (9.20) and (9.24), we find that

$$\sum_{k=1}^{n} \frac{d^k}{dt^k}\left[\tilde{\alpha}_k(t)y(t) - \tilde{\beta}_k(t)u(t)\right] = \tilde{\beta}_0(t)u(t) - \tilde{\alpha}_0(t)y(t) \tag{9.25}$$

which is equivalent to (9.3). Hence the output y of the state-space model (9.2) is also an output of the modified polynomial matrix model (9.3), where the coefficients $\tilde{\alpha}$ and $\tilde{\beta}$ are given by (9.11)-(9.12) and the states $x_1(t), \ldots, x_n(t)$ are given in terms of the input u and output y by (9.23).

9.1.2 From State-Space

The solution is given for time-invariant systems in Section 9.3.2, but I cannot find a "clean" solution for the general time-varying case.

9.2 Time-Varying Discrete-Time Models

Time-varying, discrete-time polynomial matrix models are of the form

$$\begin{aligned} & y(k+n) + \alpha_{n-1}(k)y(k+n-1) + \cdots + \alpha_0(k)y(k) \\ = {} & \beta_n(k)u(k+n) + \beta_{n-1}(k)u(k+n-1) + \cdots + \beta_0(k)u(k) \end{aligned} \tag{9.26}$$

where $y \in \mathbb{R}^p$, $u \in \mathbb{R}^m$, $\alpha_0, \ldots, \alpha_{n-1} \in \mathbb{R}^{p\times p}$, and $\beta_0, \ldots, \beta_n \in \mathbb{R}^{p\times m}$.

Remark 9.3 Discrete-time polynomial matrix models are sometimes called *time-series.*

9.2.1 To State-Space

In this section, we show how to construct a time-varying discrete-time state-space model of the form

$$\begin{aligned} x(k+1) &= A(k)x(k) + B(k)u(k) \\ y(k) &= C(k)x(k) + D(k)u(k) \end{aligned} \tag{9.27}$$

with the same output y as the time-varying discrete-time polynomial matrix model (9.26). Specifically, let y be the output of the state-space model (9.27), where

$$A(k) \triangleq \begin{bmatrix} 0_{p\times p} & \cdots & \cdots & 0_{p\times p} & -\alpha_0(k-0) \\ I_p & \ddots & & \vdots & -\alpha_1(k-1) \\ 0_{p\times p} & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0_{p\times p} & -\alpha_{n-2}(k-n+2) \\ 0_{p\times p} & \cdots & 0_{p\times p} & I_p & -\alpha_{n-1}(k-n+1) \end{bmatrix} \in \mathbb{R}^{np\times np} \tag{9.28}$$

$$B(k) \triangleq \begin{bmatrix} \beta_0(k-0) - \alpha_0(k-0)\beta_n(k-n) \\ \vdots \\ \beta_{n-1}(k-n+1) - \alpha_{n-1}(k-n+1)\beta_n(k-n) \end{bmatrix} \in \mathbb{R}^{np\times m} \tag{9.29}$$

$$C(k) \triangleq \begin{bmatrix} 0_{p\times p} & \cdots & 0_{p\times p} & I_p \end{bmatrix} \in \mathbb{R}^{p\times np} \tag{9.30}$$

$$D(k) \triangleq \beta_n(k-n) \in \mathbb{R}^{p\times m} \tag{9.31}$$

that is,

$$x(k) \triangleq \begin{bmatrix} x_1(k) \\ \vdots \\ x_n(k) \end{bmatrix} \in \mathbb{R}^{np\times 1} \tag{9.32}$$

$$x_1(k+1) \triangleq -\alpha_0(k)x_n(k) + \Big[\beta_0(k) - \alpha_0(k)\beta_n(k-n)\Big]u(k) \tag{9.33}$$

$$\begin{aligned} x_{i+1}(k+1) \triangleq {} & x_i(k) - \alpha_i(k-i)x_n(k) \\ & + \Big[\beta_i(k-i) - \alpha_i(k-i)\beta_n(k-n)\Big]u(k), \quad i = 1,\ldots,n-1 \end{aligned} \tag{9.34}$$

$$y(k) \triangleq x_n(k) + \beta_n(k-n)u(k) \tag{9.35}$$

Then we will show that y is also an output of the polynomial matrix model (9.26). Specifically, from (9.33)-(9.35), for $i = 1,\ldots,n-1$, we have that

$$x_1(k+1) = -\alpha_0(k)y(k) + \beta_0(k)u(k) \tag{9.36}$$

$$x_i(k) = x_{i+1}(k+1) + \alpha_i(k-i)y(k) - \beta_i(k-i)u(k) \tag{9.37}$$

Hence from (9.35) and (9.37), it follows that

$$\begin{aligned} x_{n-1}(k) = {} & y(k+1) - \beta_n(k-n+1)u(k+1) \\ & + \alpha_{n-1}(k-n+1)y(k) - \beta_{n-1}(k-n+1)u(k) \end{aligned} \tag{9.38}$$

Furthermore, recursively substituting into (9.37), we find that $x_i(k)$ is given by

$$x_i(k) = \sum_{j=0}^{n-i} \Big[\alpha_{i+j}(k-i)y(k+j) - \beta_{i+j}(k-i)u(k+j)\Big], \quad i = 1,\ldots,n \tag{9.39}$$

where $\alpha_n = I_p$. Therefore, evaluating (9.39) at $k+1$ and $i = 1$ yields

$$x_1(k+1) = \sum_{j=1}^{n} \Big[\alpha_j(k)y(k+j) - \beta_j(k)u(k+j)\Big] \tag{9.40}$$

Finally, comparing (9.36) and (9.40), we find that

$$\sum_{j=1}^{n} \Big[\alpha_j(k)y(k+j) - \beta_j(k)u(k+j)\Big] = -\alpha_0(k)y(k) + \beta_0(k)u(k) \tag{9.41}$$

which is equivalent to (9.26). Hence the output y of the state-space model (9.27) is also an output of the polynomial matrix model (9.26), where the states $x_1(k), \ldots, x_n(k)$ are given in terms of the input u and output y by (9.39).

9.2.2 From State-Space

The solution is given for time-invariant systems in Section 9.3.3, but I cannot find a "clean" solution for the general time-varying case.

9.3 Time-Invariant Models

Time-invariant, continuous-time polynomial matrix models are of the form

$$\begin{aligned} & y^{(n)}(t) + \alpha_{n-1} y^{(n-1)}(t) + \cdots + \alpha_0 y(t) \\ &= \beta_n u^{(n)}(t) + \beta_{n-1} u^{(n-1)}(t) + \cdots + \beta_0 u(t) \end{aligned} \tag{9.42}$$

where $y \in \mathbb{R}^p$, $u \in \mathbb{R}^m$, $\alpha_0, \ldots, \alpha_{n-1} \in \mathbb{R}^{p \times p}$, and $\beta_0, \ldots, \beta_n \in \mathbb{R}^{p \times m}$.

Time-invariant, discrete-time polynomial matrix models are of the form

$$\begin{aligned} & y(k+n) + \alpha_{n-1} y(k+n-1) + \cdots + \alpha_0 y(k) \\ &= \beta_n u(k+n) + \beta_{n-1} u(k+n-1) + \cdots + \beta_0 u(k) \end{aligned} \tag{9.43}$$

9.3.1 To State-Space

The time-invariant polynomial matrix models (9.42) and (9.43) are special cases of the time-varying forms (9.1) and (9.26). Hence letting

$$A \triangleq \begin{bmatrix} 0_{p\times p} & \cdots & \cdots & 0_{p\times p} & -\alpha_0 \\ I_p & \ddots & & \vdots & -\alpha_1 \\ 0_{p\times p} & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0_{p\times p} & -\alpha_{n-2} \\ 0_{p\times p} & \cdots & 0_{p\times p} & I_p & -\alpha_{n-1} \end{bmatrix} \in \mathbb{R}^{np \times np} \tag{9.44}$$

$$B \triangleq \begin{bmatrix} \beta_0 - \alpha_0 \beta_n \\ \vdots \\ \beta_{n-1} - \alpha_{n-1} \beta_n \end{bmatrix} \in \mathbb{R}^{np \times m} \tag{9.45}$$

$$C \triangleq \begin{bmatrix} 0_{p\times p} & \cdots & 0_{p\times p} & I_p \end{bmatrix} \in \mathbb{R}^{p \times np} \tag{9.46}$$

$$D \triangleq \beta_n \in \mathbb{R}^{p \times m} \tag{9.47}$$

then from Section 9.1.1, we have that the output y of the state-space model

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) + Du(t) \end{aligned} \tag{9.48}$$

is also an output of the polynomial matrix model (9.42). Furthermore, from Section 9.2.1, it follows that the output y of the state-space model

$$\begin{aligned} x(k+1) &= Ax(k) + Bu(k) \\ y(k) &= Cx(k) + Du(k) \end{aligned} \tag{9.49}$$

is also an output of the polynomial matrix model (9.43).

9.3.2 From State-Space (Continuous-Time)

Let y be the output of the time-invariant, continuous-time state-space model

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) + Du(t) \end{aligned} \tag{9.50}$$

where A, B, C, and D are arbitrary matrices in $\mathbb{R}^{n\times n}$, $\mathbb{R}^{n\times m}$, $\mathbb{R}^{p\times n}$, and $\mathbb{R}^{p\times m}$. Then for $i \geq 1$, we have that

$$x^{(i)}(t) = Ax^{(i-1)}(t) + Bu^{(i-1)}(t) = A^i x(t) + \sum_{j=0}^{i-1} A^{i-1-j} B u^{(j)}(t) \tag{9.51}$$

and hence

$$\begin{aligned} y^{(i)}(t) &= Cx^{(i)}(t) + Du^{(i)}(t) && (9.52) \\ &= CA^i x(t) + Du^{(i)}(t) + C\sum_{j=0}^{i-1} A^{i-1-j} B u^{(j)}(t) && (9.53) \end{aligned}$$

Next, from the Cayley-Hamilton theorem (Theorem 5.1), it follows that there exist $\alpha_0, \ldots, \alpha_{n-1} \in \mathbb{R}$ such that

$$A^n + \alpha_{n-1}A^{n-1} + \cdots + \alpha_1 A + \alpha_0 I_n = \sum_{i=0}^{n} \alpha_i A^i = 0 \tag{9.54}$$

where $\alpha_n \triangleq 1$. Hence from (9.53) and (9.54), it follows that

$$\begin{aligned} \sum_{i=0}^{n} \alpha_i y^{(i)}(t) &= C\sum_{i=0}^{n} \alpha_i A^i x(t) + \sum_{i=0}^{n} \alpha_i \left[Du^{(i)}(t) + C\sum_{j=0}^{i-1} A^{i-1-j} B u^{(j)}(t) \right] \\ &= \sum_{i=0}^{n} \alpha_i \left[Du^{(i)}(t) + C\sum_{j=0}^{i-1} A^{i-1-j} B u^{(j)}(t) \right] && (9.55) \\ &= \sum_{i=0}^{n} \left[\alpha_i D + \sum_{j=i}^{n-1} \alpha_{j+1} C A^{j-i} B \right] u^{(i)}(t) && (9.56) \end{aligned}$$

Thus y is also an output of the polynomial matrix model

$$\begin{aligned}
y^{(n)}(t) + \alpha_{n-1}y^{(n-1)}(t) + \cdots + \alpha_0 y(t)& \\
= \beta_n u^{(n)}(t) + \beta_{n-1}u^{(n-1)}(t) + \cdots + \beta_0 u(t)&
\end{aligned} \tag{9.57}$$

where the coefficients $\alpha_0, \dots, \alpha_{n-1}$ are given by the Cayley-Hamilton theorem in (9.54), and for $i = 0, \dots, n$, we have that

$$\beta_i \triangleq \alpha_i D + \sum_{j=i}^{n-1} \alpha_{j+1} C A^{j-i} B \tag{9.58}$$

9.3.3 From State-Space (Discrete-Time)

Let y be the output of the time-invariant, discrete-time state-space model

$$\begin{aligned}
x(k+1) &= Ax(k) + Bu(k) \\
y(k) &= Cx(k) + Du(k)
\end{aligned} \tag{9.59}$$

where A, B, C, and D are arbitrary matrices in $\mathbb{R}^{n\times n}$, $\mathbb{R}^{n\times m}$, $\mathbb{R}^{p\times n}$, and $\mathbb{R}^{p\times m}$. Then for $i \geq 1$, we have that

$$x(k+i) = A^i x(k) + \sum_{j=0}^{i-1} A^{i-1-j} B u(k+j) \tag{9.60}$$

and hence

$$y(k+i) = Cx(k+i) + Du(k+i) \tag{9.61}$$

$$= CA^i x(k) + Du(k+i) + C\sum_{j=0}^{i-1} A^{i-1-j} B u(k+j) \tag{9.62}$$

Next, from the Cayley-Hamilton theorem (Theorem 5.1), it follows that there exist $\alpha_0, \dots, \alpha_{n-1} \in \mathbb{R}$ such that

$$A^n + \alpha_{n-1}A^{n-1} + \cdots + \alpha_1 A + \alpha_0 I_n = \sum_{i=0}^{n} \alpha_i A^i = 0 \tag{9.63}$$

where $\alpha_n \triangleq 1$. Hence from (9.62) and (9.63), it follows that

$$\begin{aligned}\sum_{i=0}^{n} \alpha_i y(k+i) &= \sum_{i=0}^{n} \alpha_i \left[Du(k+i) + C \sum_{j=0}^{i-1} A^{i-1-j} Bu(k+j) \right] \\ &= \sum_{i=0}^{n} \left[\alpha_i D + \sum_{j=i}^{n-1} \alpha_{j+1} C A^{j-i} B \right] u(k+i) \qquad (9.64)\end{aligned}$$

Thus y is also an output of the polynomial matrix model

$$\begin{aligned} y(k+n) + \alpha_{n-1} y(k+n-1) + \cdots + \alpha_0 y(k) \\ = \beta_n u(k+n) + \beta_{n-1} u(k+n-1) + \cdots + \beta_0 u(k) \end{aligned} \qquad (9.65)$$

where the coefficients $\alpha_0, \ldots, \alpha_{n-1}$ are given by the Cayley-Hamilton theorem in (9.63), and for $i = 0, \ldots, n$, we have that

$$\beta_i \triangleq \alpha_i D + \sum_{j=i}^{n-1} \alpha_{j+1} C A^{j-i} B \qquad (9.66)$$

9.4 Coprimeness

9.5 Controllability, Reachability, Observability, Linearity

SECTION 10

Impulse Response Models

10.1 Continuous-Time Models

Time-varying, continuous-time, infinite impulse response (IIR) models are models of the form

$$y(t) = \int_{-\infty}^{0} G(t, t+\tau)u(t+\tau)d\tau \tag{10.1}$$

where $y \in \mathbb{R}^p$, $u \in \mathbb{R}^m$, and $G : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^{p \times m}$ is called the *impulse response function.*

Remark 10.1 If u is the unit impulse at time t_0, that is,

$$u(\tau) = \delta(\tau - t_0) \tag{10.2}$$

where δ denotes the Dirac delta function, then $y(t) = G(t, t_0)$. Hence $G(t, t_0)$ is literally the response of the system to an impulse. □

Similarly, time-invariant, continuous-time, infinite impulse response models are of the form

$$y(t) = \int_{-\infty}^{0} G(\tau)u(t+\tau)d\tau \tag{10.3}$$

10.1.1 From State-Space

From the variation of parameters formula (2.3) in Section 2.1, it follows that the output y of the time-varying, continuous-time state-space model

$$\begin{aligned} \dot{x}(t) &= A(t)x(t) + B(t)u(t) \\ y(t) &= C(t)x(t) + D(t)u(t) \end{aligned} \tag{10.4}$$

is given by

$$y(t) = C(t)\left[\Phi(t, t_0)x(t_0) + \int_{t_0}^{t} \Phi(t, \tau)B(\tau)u(\tau)d\tau\right] + D(t)u(t) \tag{10.5}$$

Hence if the state-space model (10.4) is asymptotically stable, then

$$y(t) = \int_{-\infty}^{0} \Big[C(t)\Phi(t, t+\tau)B(t+\tau) + \delta(\tau)D(t)\Big]u(t+\tau)d\tau \qquad (10.6)$$

Therefore, comparing (10.1) and (10.6), we find that an asymptotically stable continuous-time state-space model of the form (10.4) is equivalently represented by the infinite impulse response model (10.1), where the impulse response function is given by

$$G(t, t+\tau) \triangleq C(t)\Phi(t, t+\tau)B(t+\tau) + \delta(\tau)D(t) \qquad (10.7)$$

10.1.2 To State-Space

10.2 Discrete-Time Models

Time-varying, discrete-time, infinite impulse response models are models of the form

$$y(k) = \sum_{\kappa=0}^{\infty} G(k, k-\kappa)u(k-\kappa) \qquad (10.8)$$

Similarly, time-invariant, discrete-time, infinite impulse response models are of the form

$$y(k) = \sum_{\kappa=0}^{\infty} G(\kappa)u(k-\kappa) \qquad (10.9)$$

The infinite impulse response models (10.1), (10.3), (10.8), and (10.9) are nonparametric since an infinite number of parameters G are required for their definition. In digital systems, however, a finite number of parameters typically suffice for modeling purposes. These models are hence called *finite impulse response* (FIR) models. They will always be assumed to be discrete-time models (otherwise they would not be finite), although they can be either time-varying or time-invariant.

Time-varying, finite impulse response (FIR) models are models of the form

$$y(k) = \sum_{\kappa=0}^{n} G(k, k-\kappa)u(k-\kappa) \qquad (10.10)$$

and time-invariant, finite impulse response (FIR) models are of the form

$$y(k) = \sum_{\kappa=0}^{n} G(\kappa)u(k-\kappa) \tag{10.11}$$

Remark 10.2 The matrices G in the discrete-time models (10.8)-(10.11) are sometimes referred to as *Markov parameters* or *impulse response parameters.* □

10.2.1 From State-Space

From the variation of parameters formula (2.17) in Section 2.3, it follows that the output y of the time-varying, discrete-time state-space model

$$\begin{aligned} x(k+1) &= A(k)x(k) + B(k)u(k) \\ y(k) &= C(k)x(k) + D(k)u(k) \end{aligned} \tag{10.12}$$

is given by

$$y(k) = C(k)\left[\Phi(k,k_0)x(k_0) + \sum_{\kappa=k_0}^{k-1} \Phi(k,\kappa+1)B(\kappa)u(\kappa)\right] + D(k)u(k) \tag{10.13}$$

Hence if the state-space model (10.12) is asymptotically stable, then

$$y(k) = D(k)u(k) + \sum_{\kappa=0}^{\infty} C(k)\Phi(k,k-\kappa)B(k-\kappa-1)u(k-\kappa-1) \tag{10.14}$$

Therefore, comparing (10.8) and (10.14), we find that an asymptotically stable discrete-time state-space model of the form (10.12) is equivalently represented by the infinite impulse response model (10.8), where the impulse response function is given by

$$G(k,k-\kappa) \triangleq C(k)\Phi(k,k-\kappa)B(k-\kappa-1)u(k-\kappa-1) + \delta_1(\kappa)D(k) \tag{10.15}$$

and $\delta_{\mathbb{1}}(\kappa)$ denotes the discrete unit impulse, that is,

$$\delta_{\mathbb{1}}(\kappa) \triangleq \begin{cases} 1, & \kappa = 0 \\ 0, & \text{otherwise} \end{cases} \tag{10.16}$$

10.2.2 To State-Space

10.3 Linearity

Theorem 10.1 The impulse response models (10.1)-(10.11) are linear.

Proof Since the time-invariant forms are special cases of the time-varying forms, we only need to show that the time-varying forms are linear. We start with the time-varying, continuous-time model (10.1).

The first task is to determine what exactly the system's state is. To accomplish this, note that (10.1) can be written as

$$y(t) = \int_{0}^{t-t_0} G(t, t-\tau)u(t-\tau)d\tau + \int_{t-t_0}^{\infty} G(t, t-\tau)u(t-\tau)d\tau$$

where the first term contains *the input u for all times $t \geq t_0$* (see Definition 4.1). Hence from Definition 4.1, it follows that the state $x(t_0)$ must contain all of the information in the second term required to *uniquely determine the output y for times $t \geq t_0$*. Thus $x(t_0)$ is the entire time history of u over $(-\infty, t_0)$. Note that the impulse response function G is part of the system definition, and therefore G is not part of the state. Furthermore, in general, the state $x(t_0)$ of (10.1) will be infinite-dimensional. With this in mind, you might think of (10.1) as

$$y(t) = \int_{0}^{t-t_0} G(t, t-\tau)u(t-\tau)d\tau + f\Big(G, t, x(t_0)\Big) \tag{10.17}$$

where the state $x(t_0)$ is infinite-dimensional and represents the time-history of u for $\tau < t_0$.

Next, using this state definition, we will check whether the system (10.1) satisfies the scalability and additivity properties of Definition 7.1. Specifically, note that the output of (10.1) due to the input αu and initial condition $\alpha x(t_0)$ is given by

$$\int\limits_0^{t-t_0} G(t,t-\tau)\Big[\alpha u(t-\tau)\Big]d\tau + f\Big(G,t,\alpha x(t_0)\Big)$$

$$= \int\limits_0^{t-t_0} G(t,t-\tau)\Big[\alpha u(t-\tau)\Big]d\tau + \int\limits_{t-t_0}^{\infty} G(t,t-\tau)\Big[\alpha u(t-\tau)\Big]d\tau$$

$$= \int\limits_0^{\infty} G(t,t-\tau)\Big[\alpha u(t-\tau)\Big]d\tau = \alpha y(t)$$

Hence (10.1) satisifies the scalability property. Furthermore, letting y_i denote the output of (10.1) due to the input u_i and initial condition $x_i(t_0)$, we have that the output of (10.1) due to the input $u_1 + u_2$ and initial condition $x_1(t_0) + x_2(t_0)$ is given by

$$\int\limits_0^{t-t_0} G(t,t-\tau)\Big[u_1(t-\tau) + u_2(t-\tau)\Big]d\tau + f\Big(G,t,x_1(t_0) + x_2(t_0)\Big)$$

$$= \int\limits_0^{t-t_0} G(t,t-\tau)\Big[u_1(t-\tau) + u_2(t-\tau)\Big]d\tau$$

$$+ \int\limits_{t-t_0}^{\infty} G(t,t-\tau)\Big[u_1(t-\tau) + u_2(t-\tau)\Big]d\tau$$

$$= \int\limits_0^{\infty} G(t,t-\tau)u_1(t-\tau)d\tau + \int\limits_0^{\infty} G(t,t-\tau)u_2(t-\tau)d\tau = y_1(t) + y_2(t)$$

Hence (10.1) also satisifies the additivity property of Defintion 7.1. Therefore (10.1) is a linear model.

Finally, in a similar fashion, note that the state $x(k_0)$ for the time-varying, discrete-time, infinite impulse response model (10.8) is the time-history of u for all times before k_0. Thus following the same procedure as in the continuous-time case, we find that time-varying, discrete-time, infinite impulse response model (10.8) satisfies the scalability and additivity properties of Definition (7.1), and therefore it is also linear. □

Remark 10.3 If you are dealing with a non-causal system, meaning that the current output depends on future inputs, then the lower integration and summation bounds in (10.1)-(10.11) will be negative. However, in this case, Theorem 10.1 does not directly apply because the definition of the state in the proof no longer meets the requirements of a state. □

10.4 From State-Space

The impulse response

SECTION 11

Other Linear Models

11.1 Frequency Response Functions

Frequency response functions model how systems respond in steady-state to sinusoidal inputs. They are usually used to model linear systems, although they are sometimes also used to model systems with static nonlinearities. They are defined as follows:

Definition 11.1 Let u and y denote the input and output of a system. Then the *frequency response function* $G(2\pi\jmath f)$ of the system is the ratio of the Fourier transform of the output to the Fourier transform of the input. Specifically, we write

$$G(2\pi\jmath f) = \frac{\mathcal{F}\big[y(t)\big]}{\mathcal{F}\big[u(t)\big]} = \frac{Y(2\pi\jmath f)}{U(2\pi\jmath f)} \tag{11.1}$$

where $\mathcal{F}$ denotes the Fourier transform, $\jmath$ denotes the imaginary number, and $f \in \mathbb{R}$ denotes the frequency.

Their usefulness is mainly due to the following fact about asymptotically stable linear systems:

Fact 11.1 If $u(t) = b\cos(2\pi f t + \eta)$ is the input to an asymptotically stable linear system, then there exists $\phi \in \mathbb{R}$ and $a \in \mathbb{R}$ such that as time approaches infinity, the output y of the system approaches $y(t) = a\sin(2\pi f t + \phi)$.

Example 11.1 Let $u(t) = b\cos(2\pi f t + \eta)$ be the input to an asymptotically stable linear system, and suppose that the output y approaches $y(t) = a\sin(2\pi f t + \phi)$ as time approaches infinity. Then the frequency response function evaluated at the frequency f is given by

$$G(2\pi\jmath f) = \frac{Y(2\pi\jmath f)}{U(2\pi\jmath f)} = \frac{ae^{\jmath\phi}}{be^{\jmath\eta}} = \left(\frac{a}{b}\right) e^{\jmath(\phi-\eta)} \tag{11.2}$$

Definition 11.2 The angle $\angle G(2\pi\jmath f)$ and magnitude $\big|G(2\pi\jmath f)\big|$ of the frequency response function ($\phi - \eta$ and a/b in Example 11.1) are called the *phase* and *gain* of the system at the frequency f since they show how much the output y will shift and increase compared to the input u.

Frequency response functions are common mostly because of how common sinusoidal inputs and disturbances tend to be. For instance, AC power signals, human voices, and many other signals can be accurately modeled as either sinusoids or as the sum of sinudoids. Hence if an engineer wants to know how much interference an electronic component picks up from an ambient AC power source at 50Hz (or 60Hz), he just needs to examine the magnitude of the frequency response function at that frequency.

Remark 11.1 Transfer functions and frequency response functions perform very similar tasks: they assume that the initial conditions are zero, and model the system in the complex plane. One of the differences is that the domain of a transfer function is the entire complex plane, while the domain of a frequency response function is only the imaginary axis. However, if we examine just the imaginary axis, we will see that they are essentially identical. The key difference is in how the functions tend to be parameterized: transfer functions are generally regarded as parametric models such as (11.5). Frequency response functions, on the other hand, are usually obtained from measured data, and hence they are typically only known (or measured) at a finite number of frequencies. Thus frequency response functions are nonparametric. □

11.2 Transfer Functions

We will not be using transfer function models. However, we mention them because of their ubiquity, and because the curious reader might be curious how they fit into the group of linear models. From [9, p. 55] we have the following definition of a transfer function:

Definition 11.3 Let u and y denote the input and output of a linear, time-invariant differential equation. Then the *transfer function* representation $G(s)$ of the differential equation is the ratio of the Laplace transform of the output to the Laplace transform of the input ***under the assumption that the initial conditions are zero.*** Specifically, we write

$$G(s) = \frac{\mathcal{L}\big[y(t)\big]}{\mathcal{L}\big[u(t)\big]} = \frac{Y(s)}{U(s)} \tag{11.3}$$

where $\mathcal{L}$ denotes the Laplace transform, $s \in \mathbb{C}$ denotes the Laplace variable, and $Y(s)$ and $U(s)$ denote the Laplace transforms of y and u respectively. □

Example 11.2 Consider the SISO ordinary differential equation (**??**). Then the transfer function representation of (**??**) is given by

$$G(s) = \frac{\alpha_n s^n + \cdots + \alpha_1 s + \alpha_0}{\beta_n s^n + \cdots + \beta_1 s + \beta_0} \tag{11.4}$$

□

Although transfer functions and polynomial matrix models look similar, there are many differences:

1) A transfer function maps $\mathbb{C} \to \mathbb{C}$. If the state, input, and output of a polynomial matrix model are in $\mathbb{R}^n$, $\mathbb{R}^m$, and $\mathbb{R}^p$, respectively, then you could say that a polynomial matrix model maps

$$\mathbb{R}^n \times \left\{\mathbb{R}^m\right\}_{t \geq t_0} \longrightarrow \left\{\mathbb{R}^p\right\}_{t \geq t_0} \tag{11.5}$$

that is, a polynomial matrix model takes in the initial condition $x(t_0)$ and input u for all $t \geq t_0$, and produces the output y for all $t \geq t_0$.

2) A transfer function ignores the initial conditions, while a polynomial matrix model does not. This is because transfer functions are usually used for answering questions of stability, where the initial condition can typically be ignored.

3) Transfer functions are usually only computed for SISO systems. To extend transfer functions to MIMO systems in a complete way requires a special representation such as the Smith-McMillan form (see [5]).

Remark 11.2 Usually, when talking about transfer functions, one implicitly assumes that the underlying system is finite-dimensional, that is, the transfer function is a rational polynomial function like in (11.5). Hence transfer functions are typically regarded as parametric models with $2n$ coefficients. There are, however, linear systems that are infinite-dimensional, such as ordinary differential equations with time delays, in which case the transfer function is not a rational polynomial function. □

PART III

Models with No Noise

Here we will attempt to identify a system when there is no noise present, that is, when both the input and output are known exactly. Surprisingly, even in this case, it is not always possible to identify the system exactly, as we demonstrate with the following example:

Example 11.3 Suppose that the acceleration a of a car can be exactly modeled by the time-invariant finite impulse response model

$$a(k+2) = G(0)u(k+2) + G(1)u(k+1) + G(2)u(k) \tag{11.6}$$

where u denotes the force applied to the gas pedal. Furthermore, suppose that the input u and output a are known for $k = 1, \ldots, N$, but $G(0), G(1), G(2) \in \mathbb{R}$ are unknown, that is,

$$\begin{aligned} a(3) &= G(0)u(3) + G(1)u(2) + G(2)u(1) \\ &\vdots \\ a(N) &= G(0)u(N) + G(1)u(N-1) + G(2)u(N-2) \end{aligned} \tag{11.7}$$

where $G(0), G(1), G(2) \in \mathbb{R}$ are unknown. If $u(k)$ is the equal to some constant η, then for $k = 1, \ldots, N$,

$$a(k) = \eta\Big(G(0) + G(1) + G(2)\Big) \tag{11.8}$$

Hence if the input is a constant, then we can only determine the sum of the coefficients $G(0), G(1), G(2)$ from u and a, that is, it is impossible to determine the coefficients individually using a constant input signal. □

As demonstrated in Example 11.3, the identifiability of a model will depend on the input signal that is used to identify it. Specifically, in Example 11.3 we showed that is not possible to identify a second-order finite impulse response model using a constant input signal, that is, the constant input signal is lacking some property that is required to identify the system. This property is called *persistency.* Roughly speaking, we say that a signal is *highly persistent* if many systems can be identified using that signal as an input. The constant input signal, on the other hand, is *weakly persistent.* We will precisely quantify *how* persistent a signal is later with the concept of the *degree of persistency.*

SECTION 12

Persistency

Consider the n^{th}-order time-invariant finite impulse response model

$$y(k) = \sum_{\kappa=0}^{n} G(\kappa)u(k-\kappa) = G(0)u(k) + \cdots + G(n)u(k-n) \tag{12.1}$$

where $y \in \mathbb{R}^p$, $u \in \mathbb{R}^m$, and $G(0), \ldots, G(n) \in \mathbb{R}^{p\times m}$. If u and y are known for $k = 1, \ldots, N$, then we have that

$$Y_{n,N} = \Theta_n \Phi_{n,N} \tag{12.2}$$

where $\Phi_{n,N}$ is called the *regression matrix*, Θ_n is called the *parameter vector* (even though it can be a matrix), and $Y_{n,N}$, Θ_n, and $\Phi_{n,N}$ are given by

$$Y_{n,N} \triangleq \begin{bmatrix} y(n+1) & \cdots & y(N) \end{bmatrix} \in \mathbb{R}^{p\times(N-n)} \tag{12.3}$$

$$\Theta_n \triangleq \begin{bmatrix} G(0) & \cdots & G(n) \end{bmatrix} \in \mathbb{R}^{p\times m(n+1)} \tag{12.4}$$

$$\Phi_{n,N} \triangleq \begin{bmatrix} u(n+1) & \cdots & u(N) \\ \vdots & & \vdots \\ u(1) & \cdots & u(N-n) \end{bmatrix} \in \mathbb{R}^{m(n+1)\times(N-n)} \tag{12.5}$$

The set of equations encapsulated by (12.2) are called the *regression equations.* As we saw in Example 11.3, there does not always exist a unique solution of to the regression equations (12.2). On the other hand, there might not even exist a single solution to the regression equations. The various possibilities are summarized with the following fact (see [8, Theorem 2.6.4]):

Fact 12.1 Let $A \in \mathbb{C}^{m\times n}$, $B \in \mathbb{C}^{p\times n}$, and $\Theta \in \mathbb{C}^{p\times m}$. Then:

(i) There exist no solutions Θ of $\Theta A = B$ if and only if

$$\texttt{rank} \begin{bmatrix} A \end{bmatrix} < \texttt{rank} \begin{bmatrix} A \\ B \end{bmatrix} \tag{12.6}$$

(ii) There exists a unique solution Θ of $\Theta A = B$ if and only if

$$\texttt{rank} \begin{bmatrix} A \end{bmatrix} = \texttt{rank} \begin{bmatrix} A \\ B \end{bmatrix} = m \tag{12.7}$$

(iii) There exist infinitely many solutions Θ of $\Theta A = B$ if and only if

$$\texttt{rank}\left[\begin{array}{c} A \end{array}\right] = \texttt{rank}\left[\begin{array}{c} A \\ B \end{array}\right] < m \tag{12.8}$$

From the Fact 12.1, it is now clear what condition we need to ensure the identifiabilty of the model (12.1):

Fact 12.2 Consider the time-invariant finite impulse response model (12.1), where the input u and output y are known at times $k = 1, \ldots, N$. Furthermore, let $\Phi_{n,N}$ be given by (12.5). Then the coefficients $G(0), \ldots, G(n) \in \mathbb{R}^{p \times m}$ of (12.1) are uniquely identifiable if and only if $\Phi_{n,N}$ has full row rank, that is,

$$\texttt{rank}\big[\Phi_{n,N}\big] = m(n+1) \tag{12.9}$$

Proof This follows directly from Fact 12.1. ◻

Fact 12.2 tells us how to check whether or not we will be able to identify the coefficients of an n^{th} order FIR model uniquely from a set of given data. However, the most important observation about this fact is that, since $\Phi_{n,N}$ is a function only of u, Fact 12.2 is actually a condition on the input. This leads us to the following definition:

Definition 12.1 Let $\Phi_{n,N}$ be constructed from u as in (12.5). Then the *degree of persistency* of $u\{1, N\}$ is the smallest nonnegative integer ℓ such that $\Phi_{\ell,N}$ does not have full rank, that is,

$$\texttt{rank}\big[\Phi_{\ell,N}\big] < m(\ell+1) \tag{12.10}$$

◻

Hence rephrasing Fact 12.2 in terms of the degree of persistency, we have that:

Fact 12.3 Consider the time-invariant finite impulse response model (12.1), where $u\{1, N\}$ has a degree of persistency of ℓ. Then the coefficients $G(0), \ldots, G(n) \in \mathbb{R}^{p \times m}$ of the FIR model (12.1) are uniquely identifiable from $u\{1, N\}$ and $y\{1, N\}$ if and only if $\ell > n$.

Proof Use Definition 12.1 in Fact 12.2. ◻

Unfortunately, computing the degree of persistency can be a slightly tedious process since we need to check that $\Phi_{\ell-i,N}$ has full row rank for all $i \in [1, \ell]$. Luckily, the following fact can save us a little bit of time:

Fact 12.4 Let $\ell \geq 1$. Then $u\{1, N\}$ has a degree of persistency of ℓ if and only if $\Phi_{\ell-1,N}$ has full row rank, and $\Phi_{\ell,N}$ does not have full row rank.

Proof If $u\{1, N\}$ has a degree of persistency of ℓ, then from Definition 12.1, $\Phi_{\ell-1,N}$ has full row rank, and $\Phi_{\ell,N}$ does not have full row rank.

Next, we will show that the reverse also holds. Specifically, suppose that $\Phi_{\ell-1,N}$ has full row rank, and $\Phi_{\ell,N}$ does not have full row rank. Furthermore, note that for $i \in [2, \ell]$,

$$\Phi_{\ell-1,N} = \left[\begin{array}{ccc} u(\ell) & \cdots & u(N) \\ \vdots & & \vdots \\ u(\ell-i+2) & \cdots & u(N-i+2) \\ \hline \multicolumn{3}{c}{\Phi_{\ell-i,N-i+1}} \end{array}\right]$$

$$\Phi_{\ell-i,N} = \left[\begin{array}{c|ccc} & u(N-i+2) & \cdots & u(N) \\ \Phi_{\ell-i,N-i+1} & \vdots & & \vdots \\ & u(N-\ell+2) & \cdots & u(N-\ell+i) \end{array}\right]$$

Then since $\Phi_{\ell-1,N}$ has full row rank, $\Phi_{\ell-i,N-i+1}$ has full row rank. Hence for all $j < \ell$, $\Phi_{j,N}$ has full row rank. Therefore $u\{1, N\}$ has a degree of persistency of ℓ. □

Before we proceed to develop more facts about the degree of persistency, let's stop and reflect on what exactly the degree of persistency is buying us:

Example 12.1 Suppose I have a satellite, which, from past experience, I know can be modelled well with an FIR model of order no greater than 10. However, due to the complexity of the satellite, I can't obtain a model of the system using first principles (Newton's laws). Hence my plan is to obtain a model of the system using system identification.

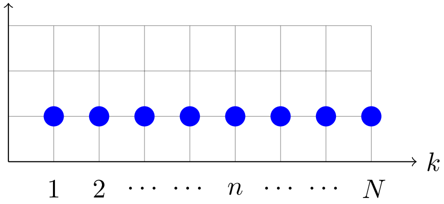
Unfortunately, due to the expensive nature of the satellite, each test is going to cost me 1 million euros, so I only want to run 1 test to determine the FIR model of the system. That means I need to be very careful about how I choose the test input. Specifically, I want to make sure that I will be able to identify the FIR model uniquely from the input/output data after I run the test, no matter what the underlying FIR model of the system is.

Fortunately, this is easily accomplished due to Fact 12.3. Specifically, from Fact 12.3, I know that if I choose a test input with a degree of persistency greater than 10, then all FIR models up to order 10 can be uniquely identified using that input. Hence before I run the test, I simply need to find a suitable input with a degree of persistency greater than 10, and use that input in the test. □

12.1 Degree Of Persistency of Some Simple Signals

Example 12.2 Consider the constant input, that is,

$$\{u(1), \ldots, u(N)\} = \{1, \ldots, 1\}$$



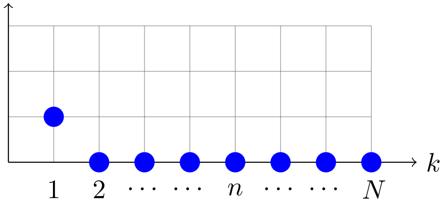
Then the regressor matrices $\Phi_{0,N}$ and $\Phi_{1,N}$ are of the form

$$\Phi_{0,N} = \left[\begin{array}{ccc} 1 & \cdots & 1 \end{array}\right], \qquad \Phi_{1,N} = \left[\begin{array}{ccc} 1 & \cdots & 1 \\ 1 & \cdots & 1 \end{array}\right]$$

and hence $\texttt{rank}\big[\Phi_{0,N}\big] = \texttt{rank}\big[\Phi_{1,N}\big] = 1$. Therefore, from Fact 12.4, $u\{1, N\}$ has a degree of persistency of 1. □

Example 12.3 Consider the unit impulse, that is,

$$\{u(1), u(2), \ldots, u(N)\} = \{1, 0, \ldots, 0\}$$



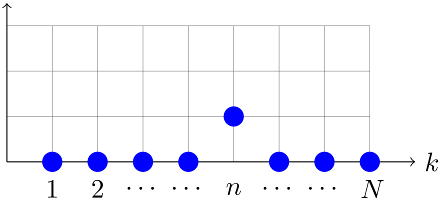
Then the regressor matrices $\Phi_{0,N}$ and $\Phi_{1,N}$ are of the form

$$\Phi_{\ell,N} = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}, \qquad \Phi_{\ell,N} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \end{bmatrix}$$

and hence $\texttt{rank}\big[\Phi_{0,N}\big] = \texttt{rank}\big[\Phi_{1,N}\big] = 1$. Therefore, from Fact 12.4, $u\{1, N\}$ has a degree of persistency of 1. □

Example 12.4 Consider the unit impulse at time $k = n$, that is,

$$\big\{u(1), \ldots, u(n-1), u(n), u(n+1), \ldots, u(N)\big\} = \big\{0, \ldots, 0, 1, 0, \ldots, 0\big\}$$



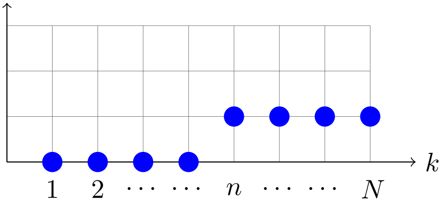
If $N \geq 2n$, then the regressor matrices $\Phi_{n-1,N}$ and $\Phi_{n,N}$ are of the form

$$\Phi_{n-1,N} = \begin{bmatrix} I_n & 0_{n\times(N-2n+1)} \end{bmatrix}, \qquad \Phi_{n,N} = \begin{bmatrix} 0_{1\times n} & 0_{1\times(N-2n)} \\ I_n & 0_{n\times(N-2n)} \end{bmatrix}$$

and hence $\texttt{rank}\big[\Phi_{n-1,N}\big] = \texttt{rank}\big[\Phi_{n,N}\big] = n$. Therefore, from Fact 12.4, $u\{1, N\}$ has a degree of persistency of n. □

Example 12.5 Consider the step input at time $k = n$, that is,

$$\{u(1), \ldots, u(n-1), u(n), \ldots, u(N)\} = \{0, \ldots, 0, 1, \ldots, 1\}$$



If $N \geq 2n$, then the regressor matrices $\Phi_{n-1,N}$ and $\Phi_{n,N}$ are of the form

$$\Phi_{n-1,N} \triangleq \left[\begin{array}{cccc|c} 1 & \cdots & \cdots & 1 & \\ 0 & \ddots & & \vdots & \\ \vdots & \ddots & \ddots & \vdots & \mathbb{1}_{n\times(N-2n+1)} \\ 0 & \cdots & 0 & 1 & \end{array}\right]$$

$$\Phi_{n,N} \triangleq \left[\begin{array}{cccc|c} \multicolumn{4}{c|}{\mathbb{1}_{1\times n}} & \mathbb{1}_{1\times(N-2n)} \\ \hline 1 & \cdots & \cdots & 1 & \\ 0 & \ddots & & \vdots & \\ \vdots & \ddots & \ddots & \vdots & \mathbb{1}_{n\times(N-2n)} \\ 0 & \cdots & 0 & 1 & \end{array}\right]$$

where $\mathbb{1}_{m\times p}$ denotes a $m \times p$ matrix of ones. Hence $\texttt{rank}\big[\Phi_{n-1,N}\big] = \texttt{rank}\big[\Phi_{n,N}\big] = n$, and therefore from Fact 12.4, $u\{1, N\}$ has a degree of persistency of n. □

12.2 The Degree of Persistency as a Polynomial Matrix Condition

Hopefully the previous examples have demonstrated why the degree of persistency is meaningful and important. Here we take a step back and analyze some alternative interpretations of the degree of persistency, specifically, in terms of polynomial matrices.

Fact 12.5 The following statements are equivalent:

(i) $u\{1, N\}$ has a degree of persistency of ℓ.

(ii) $\texttt{rank}\big[\Phi_{\ell,N}\big] < m(\ell+1)$, and either $\ell = 0$ or $\texttt{rank}\big[\Phi_{\ell-1,N}\big] = m\ell$.

(iii) ℓ is the smallest nonnegative integer for which there exist C_0, …, $C_\ell \in \mathbb{R}^{m\times m}$, where $C_0, \ldots, C_\ell$ are not all zero, such that

$$C_\ell u(k+\ell) + \cdots + C_0 u(k) = 0_{m\times 1}, \qquad \text{for all } k = 1, \ldots, N-\ell \tag{12.11}$$

(iv) ℓ is the smallest nonnegative integer for which there exists a nonzero $C \in \mathbb{R}^{m\times m}[\mathbf{q}]$ of degree less than or equal to ℓ such that

$$C(\mathbf{q})u(k) = 0_{m\times 1}, \qquad \text{for all } k = 1, \ldots, N-\ell \tag{12.12}$$

Proof From Fact 12.4, we have that (i) and (ii) are equivalent. Furthermore, letting $C(\mathbf{q}) \triangleq C_\ell \mathbf{q}^\ell + \cdots + C_1\mathbf{q} + C_0$, it follows that (iii) and (iv) are equivalent.

To show that (ii) and (iii) are equivalent, note that $\Phi_{\ell,N}$ does not have full row rank if and only if there exists a nonzero $\tilde{C} \in \mathbb{R}^{m\times m(\ell+1)}$ such that $\tilde{C}\Phi_{\ell,N} = 0$. Specifically, partitioning $\tilde{C}$ into $m\times m$ blocks, that is, $\tilde{C} = \left[\begin{array}{c|c|c} C_\ell & \cdots & C_0 \end{array}\right]$, we find that $\Phi_{\ell,N}$ does not have full row rank if and only if there exists a nonzero $\tilde{C}$ such that

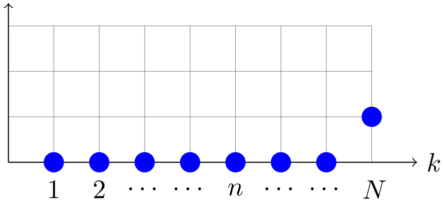
$$\begin{aligned}\tilde{C}\Phi_{\ell,N} &= \left[\begin{array}{ccc} C_\ell u(\ell+1)+\cdots+C_0 u(1), & \cdots, & C_\ell u(N)+\cdots+C_0 u(N-\ell) \end{array}\right] \\ &= \left[\begin{array}{ccc} 0_{m\times 1}, & \cdots, & 0_{m\times 1} \end{array}\right]\end{aligned}$$

that is, (12.11). Hence (ii) and (iii) are equivalent. ◻

Note that Fact 12.5 does not imply that C_ℓ in (12.11) is nonzero, that is, there can exist a nonzero $C \in \mathbb{R}^{m\times m}[\mathbf{q}]$ of degree less than ℓ which satisfies (12.12), as we demonstrate in the following example:

Example 12.6 Consider the unit impulse, where the impulse occurs at the end of the known signal, that is,

$$\{u(1), \ldots, , u(N-1), u(N)\} = \{0, \ldots, 0, 1\}$$



Then the regressor matrices $\Phi_{0,N}$ and $\Phi_{1,N}$ are of the form

$$\Phi_{0,N} \triangleq \begin{bmatrix} u(1) & \cdots & u(N-1) & u(N) \end{bmatrix} = \begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix}$$
$$\Phi_{1,N} \triangleq \begin{bmatrix} u(2) & \cdots & u(N) \\ u(1) & \cdots & u(N-1) \end{bmatrix} = \begin{bmatrix} 0_{1\times(n-2)} & 1 \\ 0_{1\times(n-2)} & 0 \end{bmatrix}$$

and hence $\texttt{rank}\big[\Phi_{0,N}\big] = \texttt{rank}\big[\Phi_{1,N}\big] = 1$. Therefore $u\{1, N\}$ has a degree of persistency of 1. Furthermore, letting $\tilde{C} = \begin{bmatrix} C_1 & C_0 \end{bmatrix} = \begin{bmatrix} 0 & 1 \end{bmatrix}$, we find that $\tilde{C}\Phi_{1,N} = 0$. Hence the polynomial $C(\mathbf{q}) \triangleq 0\mathbf{q} + 1$ is of degree 0 and satisfies (12.12). ◻

Naturally, this example leads us to wonder if there are any conditions under which we can guarantee that, if $C_0, \ldots, C_\ell \in \mathbb{R}^{m\times m}$ are not all zero and satisfy (12.11), then $C_\ell \neq 0$. One such condition is given in the following fact:

Fact 12.6 Let $u\{1, N-1\}$ and $u\{1, N\}$ have a degree of persistency of ℓ. If $C_0, \ldots, C_\ell \in \mathbb{R}^{m \times m}$ are not all zero and

$$C_\ell u(k+\ell) + \cdots + C_0 u(k) = 0_{m \times 1}, \qquad \text{for all } k = 1, \ldots, N-\ell \qquad (12.13)$$

then $C_\ell \neq 0$.

Proof Since $C_0, \ldots, C_\ell$ are not all zero, $\ell = 0$ implies that $C_\ell \neq 0$. Hence the result holds for $\ell = 0$.

Next, note that for $\ell > 0$,

$$\Phi_{\ell,N} \triangleq \begin{bmatrix} u(\ell+1) & \cdots & u(N) \\ u(\ell) & \cdots & u(N-1) \\ \vdots & & \vdots \\ u(1) & \cdots & u(N-\ell) \end{bmatrix} = \left[\begin{array}{ccc} u(\ell+1) & \cdots & u(N) \\ \hline & & \\ & \Phi_{\ell-1,N-1} & \\ & & \end{array}\right]$$

Thus letting $\bar{C} \triangleq \begin{bmatrix} C_{\ell-1} & \cdots & C_0 \end{bmatrix}$, we find that (12.13) can be written as

$$\begin{bmatrix} C_\ell & \bar{C} \end{bmatrix} \Phi_{\ell,N} = C_\ell \begin{bmatrix} u(\ell+1) & \cdots & u(N) \end{bmatrix} + \bar{C}\Phi_{\ell-1,N-1} = 0 \qquad (12.14)$$

Finally, suppose that $C_\ell = 0$. Then since $C_0, \ldots, C_\ell$ are not all zero, $\bar{C} \neq 0$. Therefore, from (12.14) it follows that $\bar{C}\Phi_{\ell-1,N-1} = 0$, that is, $\Phi_{\ell-1,N-1}$ does not have full row rank. However, this contradicts the fact that $u\{1, N-1\}$ has a degree of persistency of ℓ. Hence $C_\ell \neq 0$. □

Corollary 12.1 Let $u\{1, N-1\}$ and $u\{1, N\}$ have a degree of persistency of ℓ. If $C \in \mathbb{R}^{m \times m}[\mathbf{q}]$ is nonzero and

$$C(\mathbf{q})u(k) = 0_{m \times 1}, \qquad \text{for all } k = 1, \ldots, N-\ell \qquad (12.15)$$

then $C(\mathbf{q})$ has a degree greater than or equal to ℓ.

Proof Suppose that $C(\mathbf{q})$ has a degree less than ℓ. Then from (12.15), there exist $C_0, \ldots, C_{\ell-1} \in \mathbb{R}^{m \times m}$, where $C_0, \ldots, C_{\ell-1}$ are not all zero, such that

$$C_{\ell-1} u(k+\ell-1) + \cdots + C_0 u(k) = 0_{m \times 1}, \qquad \text{for all } k = 1, \ldots, N-\ell$$

However, this contradicts Fact 12.6. Hence $C(\mathbf{q})$ must have a degree greater than or equal to ℓ. □

Fact 12.6 and Corollary 12.1 give us conditions under which we can say something about the leading term of the annihilating polynomial $C(\mathbf{q})$. The next fact addresses the opposite problem. Specifically, if we find a $C(\mathbf{q})$ of degree ℓ which satisfies (12.15) and has an invertible leading coeffient, then the following fact reveals an interesting property of the signal $u\{1, N\}$.

Fact 12.7 Let $u\{1, N\}$ have a degree of persistency of ℓ, where $C_0, \ldots, C_\ell \in \mathbb{R}^{m \times m}$ are not all zero, and

$$C_\ell u(k+\ell) + \cdots + C_0 u(k) = 0_{m\times 1}, \qquad \text{for all } k = 1, \ldots, N-\ell \quad (12.16)$$

If C_ℓ is invertible, then $u\{1, i\}$ has a degree of persistency of ℓ for all $i \in [\ell(m+1)-1, N]$.

Proof First, note that since $u\{1, N\}$ has a degree of persistency of ℓ, $\Phi_{\ell-1,N} \in \mathbb{R}^{m\ell\times(N-\ell+1)}$ has full row rank. Hence $N \geq \ell(m+1)-1$.

Next, let $\mathcal{C} \in \mathbb{R}^{m\ell\times m\ell}$ and $U(k) \in \mathbb{R}^{m\ell}$ be given by

$$\mathcal{C} \triangleq \left[\begin{array}{ccc} -C_\ell^{-1}C_{\ell-1} & \cdots & -C_\ell^{-1}C_0 \\ \hline I_{m(\ell-1)} & & 0_{m(\ell-1)\times m} \end{array}\right], \quad U(k) = \begin{bmatrix} u(k+\ell-1) \\ \vdots \\ u(k) \end{bmatrix}$$

Then from (12.16) it follows that

$$U(k) = \mathcal{C}^{k-1}U(1), \quad \text{for all } k = 1, \ldots, N-\ell+1$$

and hence

$$\begin{aligned} \Phi_{\ell-1,N} &= \begin{bmatrix} u(\ell) & \cdots & u(N) \\ \vdots & & \vdots \\ u(1) & \cdots & u(N-\ell+1) \end{bmatrix} \\ &= \begin{bmatrix} U(1) & \cdots & U(N-\ell+1) \end{bmatrix} = \begin{bmatrix} \mathcal{C}^0 U(1) & \cdots & \mathcal{C}^{N-\ell}U(1) \end{bmatrix} \end{aligned}$$

Next, from the Cayley-Hamiton theorem (Theorem 5.1), recall that for all $i \geq m\ell$, there exist $p_0, \ldots, p_{m\ell-1} \in \mathbb{R}$, where $p_0, \ldots, p_{m\ell-1}$ are not all zero, such that

$$\mathcal{C}^i = p_0 I_{m\ell} + p_1\mathcal{C} + \cdots + +p_{m\ell-1}\mathcal{C}^{m\ell-1}$$

Hence the final columns of $\Phi_{\ell-1,N}$ are linearly dependent on the first $m\ell$ columns, that is,

$$\begin{aligned} \texttt{rank}\big[\Phi_{\ell-1,N}\big] &= \texttt{rank}\begin{bmatrix} \mathcal{C}^0 U(1) & \cdots & \mathcal{C}^{m\ell-1}U(1) \end{bmatrix} \\ &= \texttt{rank}\begin{bmatrix} U(1) & \cdots & U(m\ell) \end{bmatrix} = \texttt{rank}\big[\Phi_{\ell-1,\ell(m+1)-1}\big] \end{aligned}$$

Finally, since $u\{1, N\}$ has a degree of persistency of ℓ, it follows that $\texttt{rank}\big[\Phi_{\ell-1,N}\big] = m\ell$ and $\texttt{rank}\big[\Phi_{\ell,N}\big] < m(\ell+1)$. Therefore

$$\texttt{rank}\big[\Phi_{\ell-1,\ell(m+1)-1}\big] = m\ell \quad \text{and} \quad \texttt{rank}\big[\Phi_{\ell-1,\ell(m+1)-1}\big] < m(\ell+1)$$

Hence from Fact 12.4, $u\{1, \ell(m+1)-1\}$ has a degree of persistency of ℓ. □

12.3 Toeplitz Matrices

Next, we introduce the Toeplitz matrix $\mathcal{T}_i(C)$ for multiplying together two polynomial matrices. The following facts will make extensive use of Toeplitz matrices for making statements about solutions of the equation $D(\mathbf{q})u(k) = 0$. However, first we show how Toeplitz matrices are used to multiply polynomial matrices.

Fact 12.8 Let $C \in \mathbb{R}^{n\times m}[\mathbf{q}]$, $E \in \mathbb{R}^{p\times n}[\mathbf{q}]$, and $F \in \mathbb{R}^{p\times m}[\mathbf{q}]$, where

$$C(\mathbf{q}) \triangleq C_\ell \mathbf{q}^\ell + \cdots + C_1\mathbf{q} + C_0 \tag{12.17}$$

$$E(\mathbf{q}) \triangleq E_i \mathbf{q}^i + \cdots + E_1\mathbf{q} + E_0 \tag{12.18}$$

$$F(\mathbf{q}) \triangleq F_{\ell+i} \mathbf{q}^{\ell+i} + \cdots + F_1\mathbf{q} + F_0 \tag{12.19}$$

Furthermore, let $\mathcal{T}_i(C)$ denote the block-Toeplitz matrix

$$\mathcal{T}_i(C) \triangleq \begin{bmatrix} C_\ell & C_{\ell-1} & \cdots & C_1 & C_0 & 0 & \cdots & \cdots & 0 \\ 0 & C_\ell & C_{\ell-1} & \cdots & C_1 & C_0 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & C_\ell & C_{\ell-1} & \cdots & C_1 & C_0 & 0 \\ 0 & \cdots & \cdots & 0 & C_\ell & C_{\ell-1} & \cdots & C_1 & C_0 \end{bmatrix} \tag{12.20}$$

where $\mathcal{T}_i(C) \in \mathbb{R}^{n(i+1)\times m(\ell+i+1)}$, and the zeros denote $n \times m$ matrices of zeros. Then $F(\mathbf{q}) = E(\mathbf{q})C(\mathbf{q})$ if and only if

$$\mathcal{T}_0(F) = \mathcal{T}_0(F)\mathcal{T}_i(C) \tag{12.21}$$

Proof This is more of a derivation than a proof. Note that $\mathcal{T}_0(E) \triangleq \begin{bmatrix} E_i & \cdots & E_0 \end{bmatrix}$ and $\mathcal{T}_0(F) \triangleq \begin{bmatrix} F_{\ell+i} & \cdots & F_0 \end{bmatrix}$. Hence computing the matrix product $\mathcal{T}_0(F)\mathcal{T}_i(C)$, and comparing $\mathcal{T}_0(F)$ to the coefficients of the product $E(\mathbf{q})C(\mathbf{q})$ we find that the two forms are equivalent. □

Example 12.7 Let $C(\mathbf{q}) = C_2\mathbf{q}^2 + C_1\mathbf{q} + C_0$. Then

$$\mathcal{T}_0(C) \triangleq \begin{bmatrix} C_2 & C_1 & C_0 \end{bmatrix} \tag{12.22}$$

$$\mathcal{T}_1(C) \triangleq \begin{bmatrix} C_2 & C_1 & C_0 & 0 \\ 0 & C_2 & C_1 & C_0 \end{bmatrix} \tag{12.23}$$

$$\mathcal{T}_2(C) \triangleq \begin{bmatrix} C_2 & C_1 & C_0 & 0 & 0 \\ 0 & C_2 & C_1 & C_0 & 0 \\ 0 & 0 & C_2 & C_1 & C_0 \end{bmatrix} \tag{12.24}$$

$$\mathcal{T}_3(C) \triangleq \begin{bmatrix} C_2 & C_1 & C_0 & 0 & 0 & 0 \\ 0 & C_2 & C_1 & C_0 & 0 & 0 \\ 0 & 0 & C_2 & C_1 & C_0 & 0 \\ 0 & 0 & 0 & C_2 & C_1 & C_0 \end{bmatrix} \tag{12.25}$$

□

Next, we use the Toeplitz matrix $\mathcal{T}_i(C)$ to make a statement about solutions of the equation $D(\mathbf{q})u(k) = 0$.

Fact 12.9 Let $u\{1, N\} \in \mathbb{R}^m$ have a degree of persistency of ℓ, and let

$$C(\mathbf{q})u(k) = 0_{m\times 1}, \qquad \text{for all } k = 1, \ldots, N - \ell \tag{12.26}$$

where $C \in \mathbb{R}^{m\times m}[\mathbf{q}]$ is nonzero, and

$$C(\mathbf{q}) \triangleq C_\ell \mathbf{q}^\ell + \cdots + C_1\mathbf{q} + C_0 \tag{12.27}$$

If $i \geq 0$, $u\{1, N - i - 1\}$ has a degree of persistency of ℓ, and at least one of the following hold:

(a) $\mathcal{T}_i(C)$ has full row rank.

(b) $C(\mathbf{r})$ has full normal rank.

(c) $m = 1$

then for every $D \in \mathbb{R}^{m\times m}[\mathbf{q}]$ of degree less than or equal to $\ell + i$ which satisfies

$$D(\mathbf{q})u(k) = 0_{m\times 1}, \qquad \text{for all } k = 1, \ldots, N - \ell - i \tag{12.28}$$

there exists an $E \in \mathbb{R}^{m\times m}[\mathbf{q}]$ such that $D(\mathbf{q}) = E(\mathbf{q})C(\mathbf{q})$.

Proof First, note that if $m = 1$ and $C(\mathbf{q})$ is nonzero, then $C(\mathbf{q})$ has full row rank, that is, (c)⇒(b). Furthermore, from Fact H.2, we have that (b)⇒(a). Hence we only need to prove (a).

Next, from (12.26), note that $\mathcal{T}_i(C)\Phi_{\ell+i,N} = 0$, where

$$\Phi_{\ell+i,N} = \left[\begin{array}{ccc} u(\ell+i+1) & \cdots & u(N) \\ \vdots & & \vdots \\ u(\ell+1) & \cdots & u(N-i) \\ \hline & \Phi_{\ell-1,N-i-1} & \end{array}\right]$$

Furthermore, since $u\{1, N-i-1\}$ has a degree of persistency of ℓ, from Fact 12.5 it follows that $\Phi_{\ell-1,N-i-1}$ has full row rank. Hence

$$\texttt{rank}\big[\Phi_{\ell+i,N}\big] \geq \texttt{rank}\big[\Phi_{\ell-i,N-i-1}\big] = m\ell \tag{12.29}$$

Finally, note that the rank of a matrix plus the rank of a matrix in its nullspace must be less than or equal to its dimension. Specifically, $\texttt{rank}\big[\mathcal{T}_i(C)\big] + \texttt{rank}\big[\Phi_{\ell+i,N}\big] \leq m(\ell+i+1)$. Furthermore, since $\mathcal{T}_i(C)$ has full row rank, it follows that

$$\texttt{rank}\big[\Phi_{\ell+i,N}\big] \leq m(\ell+i+1) - m(i+1) = m\ell$$

Thus combining with (12.29), we find that $\texttt{rank}\big[\Phi_{\ell+i,N}\big] = m\ell$ and

$$\texttt{rank}\big[\mathcal{T}_i(C)\big] + \texttt{rank}\big[\Phi_{\ell+i,N}\big] = m(\ell+i+1)$$

Therefore $\mathcal{T}_i(C)$ is a complete basis for the left nullspace of $\Phi_{\ell+i,N}$, that is, for every nonzero $\tilde{D} \in \mathbb{R}^{m\times m(\ell+i+1)}$ in the left nullspace of $\Phi_{\ell+i,N}$ there exists a nonzero $\tilde{E} \in \mathbb{R}^{m\times m(i+1)}$ such that $\tilde{D} = \tilde{E}\mathcal{T}_i(C)$. Specifically, from Fact 12.8, we have that for every nonzero $D(\mathbf{q})$ which satisfies (12.28), there exists a nonzero $E \in \mathbb{R}^{m\times m}[\mathbf{q}]$ such that $D(\mathbf{q}) = E(\mathbf{q})C(\mathbf{q})$. □

At this point, you may be wondering if we really need all of the conditions in Fact 12.9. For instance, maybe it would suffice to only require that $u\{1, N-i-1\}$ has a degree of persistency of ℓ, or maybe it would suffice to only require that $\mathcal{T}_i(C)$ has full row rank. To show that both conditions are indeed required, we give counterexamples to both of these hypotheses. We start by presenting an example where $\mathcal{T}_i(C)$ has full row rank, but $u\{1, N-i-1\}$ does not have a degree of persistency of ℓ.

Example 12.8 Consider the signal

$$\{u(1), u(2), u(3), u(4)\} = \{0, 1, 0, 1\} \tag{12.30}$$

where $N = 4$, and $u\{1, 4\}$ has a degree of persistency of 2 since

$$\Phi_{0,N} = \begin{bmatrix} 0, & 1, & 0, & 1 \end{bmatrix}, \quad \Phi_{1,N} = \begin{bmatrix} 1, & 0, & 1 \\ 0, & 1, & 0 \end{bmatrix}, \quad \Phi_{2,N} = \begin{bmatrix} 0, & 1 \\ 1, & 0 \\ 0, & 1 \end{bmatrix}$$

Specifically, letting $C(\mathbf{q}) \triangleq \mathbf{q}^2 - 1$, we have that $N - \ell = 2$, and

$$C(\mathbf{q})u(k) = u(k+2) - u(k) = 0, \qquad \text{for all } k = 1, \ldots N - \ell \tag{12.31}$$

Next, let $i = 1$. Then

$$\mathcal{T}_1(C) = \begin{bmatrix} C_2 & C_1 & C_0 & 0 \\ 0 & C_2 & C_1 & C_0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \tag{12.32}$$

and thus $\mathcal{T}_1(C)$ has full row rank. However, $u\{1, N-i-1\} = u\{1, 2\}$ only has a degree of persistency of 1, and hence the conditions of Fact 12.9 are not met. Specifically, letting $D(\mathbf{q}) \triangleq 1$, we find that $N - \ell - i = 1$, and

$$D(\mathbf{q})u(k) = u(k) = 0, \qquad \text{for all } k = 1, \ldots, N - \ell - i \tag{12.33}$$

However, there does not exist an $E \in \mathbb{R}[\mathbf{q}]$ such that $D(\mathbf{q}) = E(\mathbf{q})C(\mathbf{q})$, that is,

$$\nexists E \in \mathbb{R}[\mathbf{q}] \quad \text{s.t.} \quad 1 = E(\mathbf{q})\big(\mathbf{q}^2 - 1\big) \tag{12.34}$$

□

Next, we give a case where $u\{1, N - i - 1\}$ has a degree of persistency of ℓ, but $\mathcal{T}_i(C)$ does not have full row rank.

Example 12.9 Consider the signal

$$\begin{aligned} &\{u(1), u(2), u(3), u(4), \ldots, u(20)\} \\ &\qquad = \left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \ldots, \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right\} \end{aligned} \tag{12.35}$$

where $N = 20$, and u has a degree of persistency of 1 since

$$\Phi_{0,N} \triangleq \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \end{bmatrix} \tag{12.36}$$

$$\Phi_{1,N} \triangleq \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \end{bmatrix} \tag{12.37}$$

Specifically, letting $C(\mathbf{q}) \triangleq \begin{bmatrix} 1 & -\mathbf{q} \\ 0 & 0 \end{bmatrix}$, we have that $N - \ell = 19$, and

$$C(\mathbf{q})u(k) = u_1(k) - u_2(k+1) = 0_{2\times1}, \qquad \text{for all } k = 1, \dots, N - \ell \tag{12.38}$$

where u_1 and u_2 denote the first and second rows of u, respectively.

Next, let $i \in [0, 16]$. Then $u\{1, N-i-1\}$ also has a degree of persistency of 1. However, $\mathcal{T}_i(C)$ does not have full row rank, and hence the conditions of Fact 12.9 are not met. Specifically, letting $D(\mathbf{q}) \triangleq \mathbf{q}^3 I_2$, we find that $N - \ell - i \in [3, 19]$, and

$$D(\mathbf{q})u(k) = 0_{2\times1}, \text{ for all } k = 1, \dots, N - \ell - i \tag{12.39}$$

However, there does not exist $E \in \mathbb{R}^{2\times2}[\mathbf{q}]$ such that $D(\mathbf{q}) = E(\mathbf{q})C(\mathbf{q})$, that is,

$$\nexists E \in \mathbb{R}^{2\times2}[\mathbf{q}] \quad \text{s.t.} \quad \begin{bmatrix} \mathbf{q}^3 & 0 \\ 0 & \mathbf{q}^3 \end{bmatrix} = E(\mathbf{q}) \begin{bmatrix} 1 & -\mathbf{q} \\ 0 & 0 \end{bmatrix} \tag{12.40}$$

□

12.4 Sinusoidal Persistency

Examples 12.2-12.5 demonstrated how to calculate the degree of persistency for some very simple signals, where it was easy to determine the degree of persistency from the structure of the regressor matrix. Here we calculate the degree of persistency of sinusoidal signals, encapsulated in the following fact:

Fact 12.10 Consider the sinusoidal signal

$$u(k) = \sin(k\omega + \phi) \tag{12.41}$$

where $u(1), \ldots, u(N)$ are known. Then

$$u(k+2) = 2\cos(\omega)u(k+1) - u(k), \qquad \text{for all } k = 1, \ldots, N-2 \quad (12.42)$$

Furthermore, if $N \geq 3$, and ω is not an integer multiple of 2π, then $u\{1, N\}$ has a degree of persistency of 2.

Proof (12.42) can be verified by trigonometric relations. Furthermore, from (12.42), it follows that $\Phi_{2,N}$ does not have full row rank.

Next, let $N \geq 2$. Then

$$\Phi_{1,N} = \begin{bmatrix} u(2) & u(3) & \cdots & u(N) \\ u(1) & u(2) & \cdots & u(N-1) \end{bmatrix}$$

Hence computing the determinant of the submatrix constructed from the first two columns, we find that

$$\begin{aligned} \texttt{det}\begin{bmatrix} u(2) & u(3) \\ u(1) & u(2) \end{bmatrix} &= u^2(2) - u(1)u(3) \\ &= \sin^2(2\omega + \phi) - \sin(\omega + \phi)\sin(3\omega + \phi) \\ &= \sin^2(\omega) \end{aligned}$$

Therefore, if ω is not an integer multiple of 2π, then the first two columns of $\Phi_{1,N}$ are linearly independent, that is, $\texttt{rank}\big[\Phi_{1,N}\big] = 2$.

Finally, since $\Phi_{1,N}$ has full row rank, and $\Phi_{2,N}$ does not have full row rank, then from Fact 12.4, it follows that $u\{1, N\}$ has a degree of persistency of 2. □

12.5 Persistency of the Sum

This observation leads to the following bound on the degree of persistency for the sum of two signals:

Fact 12.11 Let u and v be signals in $\mathbb{R}^m$, where $u\{1, N\}$ has a degree of persistency of ℓ_1 and $v\{1, N\}$ has a degree of persistency of ℓ_2. Also, let $C_0, \ldots, C_{\ell_1} \in \mathbb{R}^{m \times m}$ and $D_0, \ldots, D_{\ell_2} \in \mathbb{R}^{m \times m}$ satisfy

$$\begin{aligned} C_{\ell_1}u(k+\ell_1) + \cdots + C_0 u(k) &= 0_{m\times 1}, \qquad \text{for all } k = 1, \ldots, N-\ell_1 \\ D_{\ell_2}v(k+\ell_2) + \cdots + D_0 v(k) &= 0_{m\times 1}, \qquad \text{for all } k = 1, \ldots, N-\ell_2 \end{aligned} \quad (12.43)$$

that is,

$$\begin{aligned} C(\mathbf{q})u(k) &= 0_{m\times 1}, \qquad \text{for all } k = 1,\ldots,N-\ell_1 \\ D(\mathbf{q})v(k) &= 0_{m\times 1}, \qquad \text{for all } k = 1,\ldots,N-\ell_2 \\ C(\mathbf{q}) &\triangleq C_{\ell_1}\mathbf{q}^{\ell_1} + \cdots + C_1\mathbf{q} + C_0 \\ D(\mathbf{q}) &\triangleq D_{\ell_2}\mathbf{q}^{\ell_2} + \cdots + D_1\mathbf{q} + D_0 \end{aligned} \tag{12.44}$$

If $C(\mathbf{q})$ and $D(\mathbf{q})$ commute, then the degree of persistency of the sum $u\{1,N\} + v\{1,N\}$ is less than or equal to $\ell_1 + \ell_2$. Specifically,

$$D(\mathbf{q})C(\mathbf{q})\Big[u(k) + v(k)\Big] = 0_{m\times 1}, \qquad \text{for all } k = 1,\ldots,N-\ell_1-\ell_2 \tag{12.45}$$

Proof Note that

$$D(\mathbf{q})C(\mathbf{q})\big[u(k)+v(k)\big] = D(\mathbf{q})C(\mathbf{q})u(k) + D(\mathbf{q})C(\mathbf{q})v(k)$$

where, since $C(\mathbf{q})$ and $D(\mathbf{q})$ commute,

$$D(\mathbf{q})C(\mathbf{q})\big[u(k)+v(k)\big] = D(\mathbf{q})\big[C(\mathbf{q})u(k)\big] + C(\mathbf{q})\big[D(\mathbf{q})v(k)\big]$$

Hence using (12.44), we find (12.45), that is, the degree of persistency of the sum $u\{1,N\} + v\{1,N\}$ is less than or equal to $\ell_1 + \ell_2$. □

Fact 12.11 is somewhat weak, since it only gives us an upper bound. Often, a lower bound on the degree of persistency would be more useful. We could improve on the bound in (12.11) if $C(\mathbf{q})$ and $D(\mathbf{q})$ share a common right factor, however, this will just buy us a tighter upper bound. Fortunately, for scalar signals $m = 1$, we can make a more precise statement:

Fact 12.12 If all of the following hold:

(i) $u\{1,N\}$ and $u\{1,N-\ell_v-1\}$ have a degree of persistency of ℓ_u.

(ii) $v\{1,N\}$ and $v\{1,N-\ell_u-1\}$ have a degree of persistency of ℓ_v.

(iii) $D_u \in \mathbb{R}[\mathbf{q}]$ is nonzero, has a degree of ℓ_u, and

$$D_u(\mathbf{q})u(k) = 0, \qquad \text{for all } k = 1,\ldots,N-\ell_u \tag{12.46}$$

(iv) $D_v \in \mathbb{R}[\mathbf{q}]$ is nonzero, has a degree of ℓ_v, and

$$D_v(\mathbf{q})v(k) = 0, \qquad \text{for all } k = 1, \ldots, N - \ell_v \tag{12.47}$$

(v) $D_u(\mathbf{q})$ and $D_v(\mathbf{q})$ have no common roots.

(vi) $w(k) \triangleq u(k) + v(k)$

(vii) $w\{1, N\}$ and $w\{1, N-\ell_u-\ell_v-1\}$ have the same degree of persistency.

then $w\{1, N\}$ has a degree of persistency of $\ell_u + \ell_v$.

Proof First, from (iv) and (v), note that

$$D_u(\mathbf{q})D_v(\mathbf{q})w(k) = 0, \qquad \text{for all } k = 1, \ldots, N - \ell_u - \ell_v$$

Next, suppose that $w\{1, N\}$ and $w\{1, N - \ell_u - \ell_v + \ell_c - 1\}$ both have a degree of persistency of ℓ_c, where ℓ_c is less than $\ell_u + \ell_v$. Then there exists a nonzero $C \in \mathbb{R}[\mathbf{q}]$ of degree less than or equal to ℓ_c such that

$$C(\mathbf{q})w(k) = 0, \qquad \text{for all } k = 1, \ldots, N - \ell_c \tag{12.48}$$

where, from Corollary 12.1, $C(\mathbf{q})$ must have a degree of exactly ℓ_c. Furthermore, from Fact 12.9, there exists an $E \in \mathbb{R}[\mathbf{q}]$ such that

$$E(\mathbf{q})C(\mathbf{q}) = D_u(\mathbf{q})D_v(\mathbf{q})$$

Next, since $D_u(\mathbf{q})$ and $D_v(\mathbf{q})$ have no common roots, we can factor $E(\mathbf{q})$ and $C(\mathbf{q})$ into roots of $D_u(\mathbf{q})$ and $D_v(\mathbf{q})$, that is,

$$\begin{aligned} E(\mathbf{q}) &= E_u(\mathbf{q})E_v(\mathbf{q}), & D_u(\mathbf{q}) &= E_u(\mathbf{q})C_u(\mathbf{q}) \\ C(\mathbf{q}) &= C_u(\mathbf{q})C_v(\mathbf{q}), & D_v(\mathbf{q}) &= E_v(\mathbf{q})C_v(\mathbf{q}) \end{aligned}$$

Hence letting p_u, p_v, q_u, and q_v denote the degrees of $E_u(\mathbf{q})$, $E_v(\mathbf{q})$, $C_u(\mathbf{q})$, and $C_v(\mathbf{q})$, respectively, we find that

$$\ell_u = p_u + q_u, \qquad \ell_v = p_v + q_v, \qquad \ell_c = q_u + q_v$$

Next, multiplying (12.48) by $E_u(\mathbf{q})$ we find that

$$E_u(\mathbf{q})C(\mathbf{q})w(k) = C_v(\mathbf{q})D_u(\mathbf{q})w(k) = 0, \qquad \text{for all } k = 1, \ldots, N - \ell_c - p_u$$

where, from (12.46) we have that

$$E_u(\mathbf{q})C(\mathbf{q})w(k) = C_v(\mathbf{q})D_u(\mathbf{q})v(k) = 0, \qquad \text{for all } k = 1, \ldots, N - \ell_u - q_v$$

Furthermore, since $v\{1, N-\ell_u-1\}$ has a degree of persistency of ℓ_v, and $q_v \leq \ell_v$, then from (12.47) and Fact 12.9, it follows that there exists $F_v \in \mathbb{R}[\mathbf{q}]$ such that

$$F_v(\mathbf{q})D_v(\mathbf{q}) = C_v(\mathbf{q})D_u(\mathbf{q})$$

Therefore, since $D_v(\mathbf{q}) = E_v(\mathbf{q})C_v(\mathbf{q})$, it follows that $F_v(\mathbf{q})E_v(\mathbf{q}) = D_u(\mathbf{q})$, and hence $E_v(\mathbf{q})$ is a factor of both $D_u(\mathbf{q})$ and $D_v(\mathbf{q})$. However, since $D_u(\mathbf{q})$ and $D_v(\mathbf{q})$ have no common factors, $E_v(\mathbf{q})$ must be equal to a nonzero constant $e_v \in \mathbb{R}$. Similarly, we find that $E_u(\mathbf{q})$ must be equal to a nonzero constant $e_u \in \mathbb{R}$. Hence

$$C(\mathbf{q}) = \left[\tfrac{1}{e_u e_v}\right] D_u(\mathbf{q})D_v(\mathbf{q})$$

which contradicts the fact that $C(\mathbf{q})$ has a degree less than $\ell_u + \ell_v$. □

WARNING Condition (vii) makes the previous fact almost useable. From experience, it appears that this condition can be removed, but the proof is elusive.

Example 12.10 Sum of sinusoids rank example.
Show that increasing the number of sinusoids increases the rank of $\Phi_{n,N}$. However, numerically, this gets more and more poorly conditioned, so that persistency is lost. □

Example 12.11 Sum of sinusoids rank example.
Show that how the frequency content is spread out has a drastic influence on the numerical persistency. □

Example 12.12 Sum of sinusoids rank example.
Show that in general, even when the frequency content is spread out, that phasing does not help. Particularly, show that Schroeder phasing doesn't help. Furthermore, if we consider a lot of random instances of the phasing, they generally have the same persistency. □

12.6 Increasing the Degree of Persistency

It is not uncommon that instead of creating a brand new signal, we want to increase the degree of persistency of an existing signal. For instance, I might want to make sure that the underlying signal contains a sinusoid of 50Hz, but other than that, I don't care. In that case, I want to look at method for increasing the degree of persistency.

Fact 12.13 Let $u\{1,N\} \in \mathbb{R}^m$ have a degree of persistency of ℓ, where $C_0, \ldots, C_\ell \in \mathbb{R}^{m \times m}$ are not all zero, $\tilde{C} \triangleq \begin{bmatrix} C_\ell & \cdots & C_0 \end{bmatrix}$, and

$$C_\ell u(k+\ell) + \cdots + C_0 u(k) = 0_{m\times 1}, \qquad \text{for all } k = 1, \ldots, N-\ell \qquad (12.49)$$

If $u\{1, N+p\}$ has a degree of persistency of $\ell+1$, then $p \geq \mathtt{rank}\big[\tilde{C}\big]$.

Proof Since $u\{1,N\} \in \mathbb{R}^m$ has a degree of persistency of ℓ, $\Phi_{\ell,N} \in \mathbb{R}^{m(\ell+1)\times(N-n)}$ does not have full row rank. Specifically, since $\tilde{C}$ is in the left nullspace of $\Phi_{\ell,N}$, we have that

$$\mathtt{rank}\big[\Phi_{\ell,N}\big] \leq m(\ell+1) - \mathtt{rank}\big[\tilde{C}\big]$$

Next, suppose that $u\{1,N+p\}$ has a degree of persistency of $\ell+1$. Then $\Phi_{\ell,N+p} \in \mathbb{R}^{m(\ell+1)\times(N-\ell+p)}$ has full row rank, that is,

$$\mathtt{rank}\big[\Phi_{\ell,N+p}\big] = m(\ell+1)$$

Furthermore, since

$$\Phi_{\ell,N+p} = \left[\begin{array}{c|ccc} & u(N+1) & \cdots & u(N+p) \\ \Phi_{\ell,N} & \vdots & & \vdots \\ & u(N-\ell+1) & \cdots & u(N+p-\ell) \end{array}\right]$$

it follows that

$$\mathtt{rank}\big[\Phi_{\ell,N+p}\big] \leq \mathtt{rank}\big[\Phi_{\ell,N}\big] + p \leq m(\ell+1) - \mathtt{rank}\big[\tilde{C}\big] + p$$

Hence $p \geq \mathtt{rank}\big[\tilde{C}\big]$. ◻

12.6.1 Zero Buffering

Now we be a good time to mention that...

SECTION 13

Persistency in Polynomial Matrix Models

Be very careful here. This will be the basis for subspace identification!!!!!

Could you elaborate on that please....

SECTION 14

Persistency in the Differentiation Operator

Show that persistency has an almost identical interpretation when we consider models in the differentiation operator.

SECTION 15

Degree Estimation

Degree estimation is often performed by using the eigensystem realization algorithm, where the degree estimate is taken to be the rank of the Markov block Hankel matrix [10, 11]. However, this approach pre-supposes knowledge of the system's Markov parameters. Here we show that a degree estimate can be obtained directly from the regressor matrix (15.2). We begin by considering persistency in the context of the more general system (**??**). Specifically, consider the comonic system

$$A(\mathbf{r})y(k) = B(\mathbf{r})u(k), \tag{15.1}$$

where $A(\mathbf{r})$ and $B(\mathbf{r})$ are given by

$$\begin{aligned} A(\mathbf{r}) &\triangleq I_p + A_1\mathbf{r} + \cdots + A_{n-1}\mathbf{r}^{n-1} + A_n\mathbf{r}^n, \\ B(\mathbf{r}) &\triangleq B_0 + B_1\mathbf{r} + \cdots + B_{n-1}\mathbf{r}^{n-1} + B_n\mathbf{r}^n, \end{aligned}$$

and the regressor matrix Φ_N is now given by

$$\Phi_N \triangleq \begin{bmatrix} \Phi_{u,N} \\ \Phi_{y,N} \end{bmatrix}, \tag{15.2}$$

$$\Phi_{u,N} \triangleq \begin{bmatrix} u(n+1) & \cdots & u(N) \\ \vdots & & \vdots \\ u(1) & \cdots & u(N-n) \end{bmatrix} \in \mathbb{R}^{m(n+1)\times(N-n)}, \tag{15.3}$$

$$\Phi_{y,N} \triangleq \begin{bmatrix} y(n) & \cdots & y(N) \\ \vdots & & \vdots \\ y(1) & \cdots & y(N-n) \end{bmatrix} \in \mathbb{R}^{np\times(N-n)}. \tag{15.4}$$

Furthermore, the regression equation is still given by (**??**), although Θ is now given by

$$\Theta \triangleq \begin{bmatrix} B_0, & \cdots, & B_n, & -A_1, & \cdots, & -A_n \end{bmatrix}.$$

Also, note that the matrix $\Phi_{u,N}$ is the regressor matrix for the finite impulse response system (10.11). Hence, we can immediately establish some results concerning persistency for the more general model (15.1).

Proposition 15.1 If Φ_N given by (15.2) has full row rank, then $u \in \mathbb{R}^m$ has a degree of persistency ℓ which is greater than n.

Proof Since Φ_N has full row rank, both $\Phi_{y,N}$ and $\Phi_{u,N}$ have full row rank. Hence, from Theorem **??**, since $\Phi_{u,N}$ has full row rank, then $\ell > n$. □

Next, note that when the degree n of (A, B) is unknown, we cannot exactly construct the regressor matrix Φ_N in (15.2). Instead we use an estimate $\hat{n}$ of n, which may bear no resemblance to n. Specifically, let $\hat{\Phi}_N$ denote the regressor matrix (15.2) where n is replaced by $\hat{n}$. Then we have the following theorem:

Theorem 15.1 Consider the system (15.1), where (A, B) has a degree of n, and the estimate of the degree of (A, B) is $\hat{n}$. Furthermore, let $u \in \mathbb{R}^m$ have a degree of persistency of ℓ, where $C_0, \ldots, C_\ell \in \mathbb{R}^{m\times m}$ are not all zero and

$$C_0 u(k+\ell) + \cdots + C_\ell u(k) = 0_{m\times 1},$$

for all $k = 1, \ldots, N-\ell$. If C_0 is nonsingular, then

$$m \cdot \min(\ell, \hat{n}+1) \leq \mathtt{rank}\big[\hat{\Phi}_N\big] \leq m \cdot \min(\ell, \hat{n}+1) + p \cdot \min(\hat{n}, n), \quad (15.5)$$

where $\hat{\Phi}_N$ is given by (15.2) with n replaced by $\hat{n}$.

Proof First, let $\hat{\Phi}_{u,N}$ denote the matrix containing the first $m(\hat{n}+1)$ rows of $\hat{\Phi}_N$. Then $\mathtt{rank}\big[\hat{\Phi}_{u,N}\big] \leq \mathtt{rank}\big[\hat{\Phi}_N\big]$, and from Theorem **??**, $m \cdot \min(\ell, \hat{n}+1) \leq \mathtt{rank}\big[\hat{\Phi}_N\big]$.

Next, suppose that $n < \hat{n}$. Also, let

$$\begin{aligned} \vec{\phi}_{u,N} &\triangleq \begin{bmatrix} u(n+1), & \cdots, & u(N) \end{bmatrix}, \\ \vec{\phi}_{y,N} &\triangleq \begin{bmatrix} y(n+1), & \cdots, & y(N) \end{bmatrix}, \end{aligned}$$

where $\mathbf{r}^i\vec{\phi}_{u,N}$ and $\mathbf{r}^i\vec{\phi}_{y,N}$ are the i^{th} block-rows of $\hat{\Phi}_{u,N}$ and $\hat{\Phi}_{y,N}$, respectively. Then for every $i \in [0, \hat{n}-n]$, we have that

$$\mathbf{r}^i\vec{\phi}_{y,N} = -\sum_{j=1}^{n} A_j\left[\mathbf{r}^{i+j}\vec{\phi}_{y,N}\right] + \sum_{j=0}^{n} B_j\left[\mathbf{r}^{i+j}\vec{\phi}_{u,N}\right].$$

Therefore, at most pn rows of $\hat{\Phi}_{y,N}$ are linearly independent of $\hat{\Phi}_{u,N}$, and hence it follows that $\texttt{rank}\big[\hat{\Phi}_N\big] \leq \texttt{rank}\big[\hat{\Phi}_{u,N}\big] + pn$. Furthermore, since C_0 is nonsingular, then from Theorem **??**, $\texttt{rank}\big[\hat{\Phi}_N\big] \leq m \cdot \min(\ell, \hat{n}+1) + pn$.

Finally, if $\hat{n} \leq n$, then

$$\begin{aligned}\texttt{rank}\big[\hat{\Phi}_N\big] &\leq \texttt{rank}\big[\hat{\Phi}_{u,N}\big] + \texttt{rank}\big[\hat{\Phi}_{y,N}\big],\\ &= m \cdot \min(\ell, \hat{n}+1) + \texttt{rank}\big[\hat{\Phi}_{y,N}\big],\\ &\leq m \cdot \min(\ell, \hat{n}+1) + p\hat{n}.\end{aligned}$$

□

Numerical testing suggests that for SISO systems,

$$\texttt{rank}\big[\hat{\Phi}_N\big] = \min(\ell, \hat{n}+1) + \min(\hat{n}, n) \tag{15.6}$$

for almost all initial conditions of y. However, the following example demonstrates a specific case in which $\texttt{rank}\big[\hat{\Phi}_N\big] < \min(\ell, \hat{n}+1) + \min(\hat{n}, n)$.

Example 15.1 Consider the SISO system

$$(1 - a\mathbf{q})\, y(k) = u(k), \tag{15.7}$$

where $u(k) = q^k$, and $q \neq 0$. Then $n = 1$, and since u satisfies $u(k+1) = qu(k)$, its degree of persistency ℓ is 1. Hence letting $\hat{n} \geq 1$ and

$$\alpha \triangleq \frac{q^2}{q-a}, \quad y(1) \triangleq \alpha q, \tag{15.8}$$

it follows that $y(k) = \alpha u(k)$. Therefore $\texttt{rank}\big[\hat{\Phi}_N\big] = 1 < \min(\hat{n}, n) + \min(\hat{n}+1, \ell)$. However, for all other values of $y(1)$, $\texttt{rank}\big[\hat{\Phi}_N\big] = 2 = \min(\hat{n}, n) + \min(\hat{n}+1, \ell)$. □

The usefulness of Theorem 15.1 is due to the fact that the degree of persistency ℓ of the input can be computed separately from the rank of the regressor matrix $\hat{\Phi}_N$. Hence when $\hat{n} > n$ and the rank equality holds, that is, $\texttt{rank}\big[\hat{\Phi}_N\big] = m \cdot \min(\ell, \hat{n}+1) + p \cdot \min(\hat{n}, n)$, then

$$n = \frac{1}{p}\Big[\texttt{rank}\big[\hat{\Phi}_N\big] - m \cdot \min(\ell, \hat{n}+1)\Big]. \tag{15.9}$$

The following examples demonstrate this technique.

Example 15.2 Consider the linearized longitudinal model of the T-2 aircraft (**??**) given by

$$\left(1 - 1.862\mathbf{r} + 0.8798\mathbf{r}^2\right) y(k) = \left(-0.009767\mathbf{r} - 0.006026\mathbf{r}^2\right) u(k), \quad (15.10)$$

where $k \geq 1$, $y(2) = y(1) = 0$, and

$$u(k) = \cos(k/10), \quad k = 1, \ldots, 1000. \quad (15.11)$$

Also, let $\hat{n} = 6$. Then $n = 2$, $\ell = 2$, and from (15.5), we expect

$$\begin{aligned} \texttt{rank}\big[\hat{\Phi}_N\big] &= \min(\hat{n}, n) + \min(\hat{n} + 1, \ell) \\ &= 2 + 2 = 4. \end{aligned}$$

Finally, Figure 9 displays the normalized singular values ($\bar{\sigma}_i \triangleq \sigma_i/\sigma_{\max}$) of the regressor matrix $\hat{\Phi}_N$. From Figure 9, we can see that indeed $\texttt{rank}\big[\hat{\Phi}_N\big] = 4$. □

Figure 9: Normalized singular values of the regressor matrix $\hat{\Phi}_N$ for the system (15.10), where u is given by (15.11) and $\hat{n} = 6$.

Example 15.3 Consider the linearized longitudinal model of the T-2 aircraft (**??**) given by (15.10), where $k \geq 1$, $y(2) = y(1) = 0$, and

$$u(k) = \begin{cases} 0, & k = 1, \ldots, 10, \\ \cos\left([k - 11]/10\right), & 11 \leq k \leq 1000. \end{cases} \quad (15.12)$$

Also, let $\hat{n} = 6$. Then $n = 2$, $\ell = 12$, and from (15.5), we expect

$$\begin{aligned} \texttt{rank}\big[\hat{\Phi}_N\big] &= \min(\hat{n}, n) + \min(\hat{n} + 1, \ell) \\ &= 2 + 7 = 9, \end{aligned}$$

Finally, Figure 10 displays the normalized singular values ($\bar{\sigma}_i \triangleq \sigma_i/\sigma_{\max}$) of the regressor matrix $\hat{\Phi}_N$. From Figure 10, we can see that indeed $\texttt{rank}\big[\hat{\Phi}_N\big] = 9$. □

Figure 10: Normalized singular values of the regressor matrix $\hat{\Phi}_N$ for the system (15.10), where u is given by (15.12) and $\hat{n} = 6$.

PART IV

Least-Squares

SECTION 16

Sensor Noise

Suppose that we have a sensor which produced N noisy measurements y of a constant value γ, where the measurements are corrupted by the additive measurement noise w, that is,

$$y(k) = \gamma + w(k), \quad \text{for all } k = 1, \ldots, N \tag{16.1}$$

Furthermore, suppose that the noise sequence is:

$$\begin{aligned} \text{zero-mean: } & \mathbb{E}\big[w(k)\big] = 0 \text{ for all } k = 1, \ldots, N, \\ \text{white: } & \mathbb{E}\big[w(j)w(k)\big] = 0 \text{ for all } j, k = 1, \ldots, N, \text{ where } j \neq k, \\ \text{and stationary: } & \text{with variance } \sigma^2 \triangleq \mathbb{E}\big[w^2(k)\big] \text{ for all } k = 1, \ldots, N, \end{aligned}$$

where γ and σ^2 are unknown, and $\mathbb{E}$ denotes the expected value.[1] Then a natural question we might ask is:

How can we estimate γ from the measurements $y(1), \ldots, y(N)$?

Specifically, since our estimate $\hat{\gamma}_N$ of γ is a function of $y(1), \ldots, y(N)$, we could think of our estimate as a function of the data, that is,

$$\hat{\gamma}_N = f\big(y(1), \ldots, y(N)\big) \tag{16.2}$$

in which case, the question becomes:

How can we choose the function $f(\cdot)$ so that $\hat{\gamma}_N \approx \gamma$?

16.1 Sample Mean

One of the most common methods for estimating a constant value such as γ is to use the *sample mean* estimate

$$\hat{\gamma}_N = \frac{1}{N} \sum_{k=1}^{N} y(k) \tag{16.3}$$

But how *good* is this estimate?

To answer that question, first note that:

[1] We will introduce the expected value in a later chapter via Definition 23.6. For now, you might just think of $\mathbb{E}[\cdot]$ as the average value of $(\cdot)$.

- $w(1), \ldots, w(N)$ are random variables
- $y(1), \ldots, y(N)$ functions of $w(1), \ldots, w(N)$
- $\hat{\gamma}_N$ is a function of $y(1), \ldots, y(N)$

Hence the estimate $\hat{\gamma}_N$ is itself a random variable. For instance,
which has the nice property that it is an unbiased estimate of γ, that is,

$$\mathbb{E}\Big[\hat{\gamma}_N\Big] = \frac{1}{N}\sum_{k=1}^{N}\mathbb{E}\Big[\gamma + w(k)\Big] = \frac{1}{N}\sum_{k=1}^{N}\gamma = \gamma \tag{16.4}$$

Furthermore, the variance of the sample mean $\hat{\gamma}_N$ is inversely proportional to N, that is,

$$\begin{aligned}\mathbb{E}\left[\left(\hat{\gamma}_N - \gamma\right)^2\right] &= \mathbb{E}\left[\left(\left[\frac{1}{N}\sum_{k=1}^{N} y(k)\right] - \gamma\right)^2\right] \\ &= \frac{1}{N^2}\mathbb{E}\left[\left(\sum_{k=1}^{N} w(k)\right)^2\right] = \frac{\sigma^2}{N}\end{aligned} \tag{16.5}$$

Hence, the estimate $\hat{\gamma}$ will get better and better as more data is gathered, that is, as N increases.

16.2 Sample Variance

Next, note that since $y(k) - \gamma = w(k)$, it follows that $y(k) - \hat{\gamma}_N \approx w(k)$. Hence letting

$$\hat{m}(\alpha) \triangleq \frac{\alpha}{N}\sum_{k=1}^{N}\Big(y(k) - \hat{\gamma}_N\Big)^2 \tag{16.6}$$

we expect that $\big(y(k) - \hat{\gamma}_N\big)^2 \approx w^2(k)$ and $\mathbb{E}\big[\hat{m}(1)\big] \approx \sigma^2$. Specifically, the exact expected value of (16.6) is given by

$$\begin{aligned}
\mathbb{E}\Big[\hat{m}(\alpha)\Big] &= \frac{\alpha}{N}\sum_{k=1}^{N} \mathbb{E}\left[\left(y(k) - \hat{\gamma}_N\right)^2\right] && (16.7)\\
&= \frac{\alpha}{N}\sum_{k=1}^{N} \mathbb{E}\left[\left[\big(y(k) - \gamma\big) - \big(\hat{\gamma}_N - \gamma\big)\right)^2\right] && (16.8)\\
&= \frac{\alpha}{N}\sum_{k=1}^{N} \mathbb{E}\left[\Big(w(k) - \big(\hat{\gamma}_N - \gamma\big)\Big)^2\right] && (16.9)\\
&= \frac{\alpha}{N}\sum_{k=1}^{N} \mathbb{E}\Big[w^2(k) + \big(\hat{\gamma}_N - \gamma\big)^2 - 2w(k)\big(\hat{\gamma}_N - \gamma\big)\Big] && (16.10)
\end{aligned}$$

where, using (16.5), we find that

$$\begin{aligned}
\mathbb{E}\Big[\hat{m}(\alpha)\Big] &= \frac{\alpha}{N}\sum_{k=1}^{N} \left[\sigma^2 + \frac{\sigma^2}{N} - 2\mathbb{E}\Big[w(k)\big(\hat{\gamma}_N - \gamma\big)\Big]\right] && (16.11)\\
&= \frac{\alpha}{N}\sum_{k=1}^{N} \left[\sigma^2 + \frac{\sigma^2}{N} - 2\mathbb{E}\Big[w(k)\hat{\gamma}_N\Big]\right] && (16.12)
\end{aligned}$$

Furthermore, since w is a white noise sequence, the expected value of $w(k)\hat{\gamma}_N$ is given by

$$\begin{aligned}
\mathbb{E}\Big[w(k)\hat{\gamma}_N\Big] &= \mathbb{E}\left[w(k)\frac{1}{N}\sum_{k=1}^{N} y(k)\right] = \mathbb{E}\left[w(k)\frac{1}{N}\sum_{k=1}^{N} \left[\gamma + w(k)\right]\right]\\
&= \mathbb{E}\left[w(k)\frac{1}{N}\sum_{k=1}^{N} w(k)\right] = \frac{\sigma^2}{N} && (16.13)
\end{aligned}$$

and therefore

$$\mathbb{E}\Big[\hat{m}(\alpha)\Big] = \frac{\alpha}{N}\sum_{k=1}^{N} \left[\sigma^2 + \frac{\sigma^2}{N} - 2\frac{\sigma^2}{N}\right] = \frac{(N-1)\alpha\sigma^2}{N} \qquad (16.14)$$

Finally, let's go back to our initial intuition that $\mathbb{E}\big[\hat{m}(1)\big] \approx \sigma^2$. In fact, we find that

$$\mathbb{E}\Big[\hat{m}\left(1\right)\Big] = \left(\frac{N-1}{N}\right)\sigma^2 \qquad (16.15)$$

which turns out to be a pretty good estimate of σ^2 for large values of N. However, a better estimate is clearly given by $m\left(\frac{N}{N-1}\right)$. Specifically,

$$\widehat{\sigma^2} = \frac{1}{N-1} \sum_{k=1}^{N} \Big(y(k) - \hat{\gamma}_N \Big)^2 \tag{16.16}$$

is called the *sample variance*, where from (16.14), we find that the sample variance (16.16) is an unbiased estimate of σ^2, that is,

$$\mathbb{E}\left[\widehat{\sigma^2}\right] = \sigma^2 \tag{16.17}$$

16.3 Sensors with Drift

Unfortunately, (16.1) is too simplified for many practical sensors. A more accurate model is one that includes *drift*, that is, a model of the form

$$\begin{aligned} & & y(k) &= \gamma + k\beta + w(k) \\ \text{for all } j, k = 1, \ldots, N, & & \mathbb{E}\big[w(k)\big] &= 0 \\ \text{where } j \neq k: & & \mathbb{E}\big[w^2(k)\big] &= \sigma^2 \\ & & \mathbb{E}\big[w(j)w(k)\big] &= 0 \end{aligned} \tag{16.18}$$

where β denotes an unknown constant which accounts for drift in the sensor measurements. Next, we will see if we can use the same procedure as in the previous section for estimating the variance σ^2.

Let's start by examining the expected value of the sample mean (16.3) when applied to our new model (16.18), in which case, we find that

$$\begin{aligned} \mathbb{E}\left[\frac{1}{N}\sum_{k=1}^{N} y(k)\right] &= \frac{1}{N}\sum_{k=1}^{N} \mathbb{E}\Big[\gamma + k\beta + w(k)\Big] = \frac{1}{N}\sum_{k=1}^{N}\Big[\gamma + k\beta\Big] \\ &= \gamma + \left(\tfrac{N+1}{2}\right)\beta \end{aligned} \tag{16.19}$$

Is this what we want?

What we ultimately need is to compute an approximation $\hat{w}(k)$ of $w(k)$, which we can then average to compute an estimate of σ^2. Hence we need either an estimate of the sum $\gamma + k\beta$, or individual estimates of γ and β, from which we can estimate the sum $\gamma + k\beta$. Equation (16.19) provides us with none of those. However, perhaps we just considered the "wrong" sample mean. For instance, what if we scale the sample mean by time, in which case,

$$\mathbb{E}\left[\frac{1}{N}\sum_{k=1}^{N} ky(k)\right] = \frac{1}{N}\sum_{k=1}^{N}\mathbb{E}\Big[k\gamma + k^2\beta + kw(k)\Big] = \frac{1}{N}\sum_{k=1}^{N}\Big[k\gamma + k^2\beta\Big]$$

$$= \left(\tfrac{N+1}{2}\right)\gamma + \left(\tfrac{(N+1)(2N+1)}{6}\right)\beta \tag{16.20}$$

Unfortunately, this is not directly helpful either. However, there is good news: together, (16.19) and (16.20) form a system of equations from which we can estimate γ and β. Specifically, let $\hat{\gamma}$ and $\hat{\beta}$ be solutions of

$$\hat{\gamma} + \left(\tfrac{N+1}{2}\right)\hat{\beta} = \frac{1}{N}\sum_{k=1}^{N} y(k) \tag{16.21}$$

$$\left(\tfrac{N+1}{2}\right)\hat{\gamma} + \left(\tfrac{(N+1)(2N+1)}{6}\right)\hat{\beta} = \frac{1}{N}\sum_{k=1}^{N} ky(k) \tag{16.22}$$

that is, let $\hat{\gamma}$ and $\hat{\beta}$ be given by

$$\hat{\gamma} = \frac{2}{N(N-1)}\left[-3\left(\sum_{k=1}^{N} ky(k)\right) + \Big(2N+1\Big)\left(\sum_{k=1}^{N} y(k)\right)\right] \tag{16.23}$$

$$\hat{\beta} = \frac{2}{N(N-1)}\left[-3\left(\sum_{k=1}^{N} y(k)\right) + \frac{6}{N+1}\left(\sum_{k=1}^{N} ky(k)\right)\right] \tag{16.24}$$

Then $\hat{\gamma}$ and $\hat{\beta}$ are unbiased estimates of γ and β, that is,

$$\mathbb{E}\Big[\hat{\gamma}\Big] = \gamma, \qquad \mathbb{E}\Big[\hat{\beta}\Big] = \beta \tag{16.25}$$

Unfortunately, the sample variance is much more tedious to compute. We will come back to its calculation in a later chapter. However, before we do, let's step back and analyze what we are doing a little bit more systematically.

16.4 Estimating the Drift with Least-Squares

In Section 16.1, we introduced the sample mean formula (16.3). However, despite this seemingly arbitrary choice, we showed that the sample mean

(16.3) is unbiased with a variance inversely proportional to the number of measurements N. Then, in Section 16.3, we put together two unbiased estimates of γ and β using some seemingly ad-hoc variations of the sample mean formula. Pretty good for some arbitrary estimates, right?

As you probably suspected, these estimates are not arbitrary at all; they are least-squares estimates. Specifically, letting $y(k)$ be given by (16.1), we find that the sample mean (16.3) minimizes the least-squares cost function

$$J\left(\hat{\gamma}\right) = \sum_{k=1}^{N} \left[\hat{\gamma} - y(k)\right]^2 \tag{16.26}$$

since it satisfies the first-order necessary condition of optimality

$$\frac{\partial J\left(\hat{\gamma}\right)}{\partial \hat{\gamma}} = 2\sum_{k=1}^{N} \left[\hat{\gamma} - y(k)\right] = 0 \tag{16.27}$$

Similarly, letting $y(k)$ be given by (16.18), we find that the coefficients (16.23) and (16.24) minimize the least-squares cost function

$$J\left(\hat{\gamma}, \hat{\beta}\right) = \sum_{k=1}^{N} \left[\hat{\gamma} + k\hat{\beta} - y(k)\right]^2 \tag{16.28}$$

since they satisfy the first-order necessary conditions of optimality

$$\begin{align} \frac{\partial J}{\partial \hat{\gamma}} &= 2\sum_{k=1}^{N} \left[\hat{\gamma} + k\hat{\beta} - y(k)\right] \tag{16.29}\\ &= 2N\hat{\gamma} + N(N+1)\hat{\beta} - 2\left(\sum_{k=1}^{N} y(k)\right) = 0 \tag{16.30}\\ \frac{\partial J}{\partial \hat{\beta}} &= 2\sum_{k=1}^{N} k\left[\hat{\gamma} + k\hat{\beta} - y(k)\right] \tag{16.31}\\ &= N(N+1)\hat{\gamma} + \frac{1}{3}N(N+1)(2N+1)\hat{\beta} - 2\left(\sum_{k=1}^{N} ky(k)\right) = 0 \tag{16.32} \end{align}$$

Specifically, solving (16.30) and (16.32) for $\hat{\gamma}$ and $\hat{\beta}$, we find (16.23) and (16.24).

SECTION 17

The Least-Squares Estimate

Many engineering problems boil down to solving a linear system of equations $\Theta A = B$ for an unknown matrix (or vector) Θ, such as when you estimate a system with no noise present. In this case, assuming that the system was modelled correctly, there always exists at least one solution of $\Theta A = B$, and hence the main difficultly is in ensuring that Θ is uniquely identifiable.

However, when there is noise present, there typically does not exist a solution Θ of $\Theta A = B$. For instance, in the case of a noisy sensor we had the model

$$y(k) = \gamma + w(k) \tag{17.1}$$

where only $y(k)$ is known. Hence letting $\Theta \triangleq \gamma$, $A \triangleq 1$, and $B \triangleq y(k)$, it follows that there only exists a solution of $\Theta A = B$ if $w(k) = 0$.

Clearly this noisy case is quite a different problem than the one we have encountered in the previous chapters since, instead of having many possible solutions, we will typically have no solutions. However, the data *is* the data, and we can't do anything about it. Hence if there does not exist a solution to $\Theta A = B$, then we have to move on to the next best thing: a solution $\hat{\Theta}$ which *approximately* satisfies $\Theta A = B$, that is, $\hat{\Theta} A \approx B$.

There are an infinite number of approximate solutions that we could choose from, but the most popular solution is the least-squares estimate $\hat{\Theta}$ of Θ, which is the solution of

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmin}} \left\| \Theta A - B \right\|_F \tag{17.2}$$

where $\|\cdot\|_F$ denotes the *Frobenius norm* of $(\cdot)$.[2] This estimate is popular because it tends to have pretty good performance properties, and it is easy to compute. Specifically, we have the following fact:

Fact 17.1 Let $A \in \mathbb{C}^{m\times n}$, $B \in \mathbb{C}^{p\times n}$, and $\Theta \in \mathbb{C}^{p\times m}$. If A has full row rank or the product AA^H is invertible, then there exists a unique

[2] If x is a vector, that is, $x \in \mathbb{C}^p$, then the Frobenius norm and the 2-norm of x are the same, that is, $\|x\|_F = \|x\|_2$. If x is a matrix, then the Frobenius norm of x is the 2-norm of its singular values, that is, $\left\|x\right\|_F = \sqrt{\sum_i \sigma_i^2}$

least-squares solution $\hat{\Theta}$ of (17.9). Specifically, $\hat{\Theta}$ is given by

$$\hat{\Theta} = BA^T \left(AA^T\right)^{-1} \tag{17.3}$$

Proof First, note that A has full row rank if and only if AA^T is invertible, that is, the statements are equivalent.

Next, note that

$$\left\|\Theta A - B\right\|_F = \sqrt{\mathtt{tr}\left[(\Theta A - B)^H (\Theta A - B)\right]}$$

Hence from the first-order necessary conditions of optimality, we have that

$$\frac{\partial\|\hat{\Theta} A - B\|_F}{\partial\hat{\Theta}} = -\left(\frac{1}{2\|\hat{\Theta} A - B\|_F}\right)\frac{\partial\mathtt{tr}\left[(\Theta A - B)^H (\Theta A - B)\right]}{\partial\hat{\Theta}}$$

and from Fact **??**, it follows that

□

Remark 17.1 If A is real, then $A^H = A^T$. □

17.1 Equation Error Model

The ordinary least-squares estimate is particularly well suited for estimating the coefficients Θ of an *equation error model*, that is, a model of the form

$$y(k) = \Theta\phi(k) + w(k) \tag{17.4}$$

where $y(k) \in \mathbb{R}^p$, $\Theta \in \mathbb{R}^{p\times n}$, $\phi(k) \in \mathbb{R}^n$, $w(k) \in \mathbb{R}^p$, and w is white noise. Specifically, if $\phi(k)$ is known and $y(k)$ is sampled for $k = 1, \ldots, N$, then letting

$$Y \triangleq \left[\begin{array}{ccc} y(1) & \cdots & y(N) \end{array}\right] \tag{17.5}$$

$$\Phi \triangleq \left[\begin{array}{ccc} \phi(1) & \cdots & \phi(N) \end{array}\right] \tag{17.6}$$

$$W \triangleq \left[\begin{array}{ccc} w(1) & \cdots & w(N) \end{array}\right] \tag{17.7}$$

it follows that the regression equations are given by

$$Y = \Theta\Phi + W \tag{17.8}$$

where W is unknown since it contains the unknown noise signal. Hence the least-squares estimate $\hat{\Theta}$ of Θ is given by

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmin}} \left\| \Theta\Phi - Y \right\|_F = Y\Phi^T \left(\Phi\Phi^T\right)^{-1} \tag{17.9}$$

Specifically, combining with (17.8), we find that

$$\hat{\Theta} = \left(\Theta\Phi + W\right)\Phi^T \left(\Phi\Phi^T\right)^{-1} = \Theta + W\Phi^T \left(\Phi\Phi^T\right)^{-1} \tag{17.10}$$

Next, note that since w is a white noise signal, it follows that $\mathbb{E}\big[w(k)\big] = 0$ for all $k = 1, \ldots, N$. Hence $\mathbb{E}\left[W\right] = 0$. Furthermore, if Φ contains only deterministic values, then

$$\mathbb{E}\big[\hat{\Theta}\big] = \Theta + \mathbb{E}\big[W\big]\Phi^T \left(\Phi\Phi^T\right)^{-1} = \Theta \tag{17.11}$$

that is, if w is white noise, and the regressor matrix Φ is deterministic, then the least-squares estimate $\hat{\Theta}$ of the equation error model (17.8) is unbiased.

Next, we examine the covariance properties of the least-squares estimate. Specifically, note that

$$\hat{\Theta} - \mathbb{E}\big[\hat{\Theta}\big] = W\Phi^T \left(\Phi\Phi^T\right)^{-1} \tag{17.12}$$

and hence the covariance matrix of $\hat{\Theta}$ is given by

$$\begin{aligned}
&\mathbb{E}\left[\left(\hat{\Theta} - \mathbb{E}\big[\hat{\Theta}\big]\right)^T \left(\hat{\Theta} - \mathbb{E}\big[\hat{\Theta}\big]\right)\right] \\
&\qquad = \mathbb{E}\left[\left(\Phi\Phi^T\right)^{-1} \Phi W^T W \Phi^T \left(\Phi\Phi^T\right)^{-1}\right] \qquad (17.13) \\
&\qquad = \left(\Phi\Phi^T\right)^{-1} \Phi \mathbb{E}\left[W^T W\right] \Phi^T \left(\Phi\Phi^T\right)^{-1} \qquad (17.14)
\end{aligned}$$

where, since w is white noise,

$$\mathbb{E}\left[W^T W\right] = \sigma^2 I \tag{17.15}$$

and hence

$$\mathbb{E}\left[\left(\hat{\Theta} - \mathbb{E}\big[\hat{\Theta}\big]\right)^T \left(\hat{\Theta} - \mathbb{E}\big[\hat{\Theta}\big]\right)\right] = \sigma^2 \left(\Phi\Phi^T\right)^{-1} \tag{17.16}$$

Finally, let's try to obtain an estimate of the variance σ^2. To accomplish this, note that

$$Y - \hat{\Theta}\Phi = \Theta\Phi + W - \left(\Theta + W\Phi^T \left(\Phi\Phi^T\right)^{-1}\right)\Phi \tag{17.17}$$

$$= W\left(I - \Phi^T \left(\Phi\Phi^T\right)^{-1}\Phi\right) \tag{17.18}$$

where the matrix

$$\tilde{\Phi} \triangleq I - \Phi^T \left(\Phi\Phi^T\right)^{-1}\Phi \tag{17.19}$$

has the very special properties. Specifically, $\tilde{\Phi}$ is called an *idempotent* matrix since $\tilde{\Phi}^2 = \tilde{\Phi}$. Therefore, since $\tilde{\Phi}^T = \tilde{\Phi}$ it follows that

$$\left(Y - \hat{\Theta}\Phi\right)\left(Y - \hat{\Theta}\Phi\right)^T = W\left(I - \Phi^T \left(\Phi\Phi^T\right)^{-1}\Phi\right)W^T \tag{17.20}$$

$$= \mathtt{tr}\left[W^T W\left(I - \Phi^T \left(\Phi\Phi^T\right)^{-1}\Phi\right)\right] \tag{17.21}$$

Furthermore, since the matrix Φ is deterministic, it follows that

$$\mathbb{E}\left[\left(Y - \hat{\Theta}\Phi\right)\left(Y - \hat{\Theta}\Phi\right)^T\right] = \mathtt{tr}\left[\mathbb{E}\left[W^T W\right]\left(I - \Phi^T \left(\Phi\Phi^T\right)^{-1}\Phi\right)\right]$$

$$= N\sigma^2\mathtt{tr}\left[I - \Phi^T \left(\Phi\Phi^T\right)^{-1}\Phi\right] \tag{17.22}$$

Finally, noting that idempotent matrices must have eigenvalues of either 1 or 0, it follows that

$$\mathtt{tr}\left[I - \Phi^T \left(\Phi\Phi^T\right)^{-1}\Phi\right] = \mathtt{rank}\left[I - \Phi^T \left(\Phi\Phi^T\right)^{-1}\Phi\right] \tag{17.23}$$

which, when Φ has full row rank, is given by

$$\mathtt{tr}\left[I - \Phi^T \left(\Phi\Phi^T\right)^{-1}\Phi\right] = N - n \tag{17.24}$$

Hence letting

$$\eta \triangleq \frac{1}{N\mathtt{tr}\left[I - \Phi^T \left(\Phi\Phi^T\right)^{-1}\Phi\right]} \tag{17.25}$$

we find that

$$\mathbb{E}\left[\eta\left(Y - \hat{\Theta}\Phi\right)\left(Y - \hat{\Theta}\Phi\right)^T\right] = \sigma^2$$

that is,

$$\eta \left(Y - \hat{\Theta}\Phi\right) \left(Y - \hat{\Theta}\Phi\right)^T \tag{17.26}$$

is an unbiased estimator for σ^2.

Actually, we don't need $\hat{\Theta}$ at all since

$$Y - \hat{\Theta}\Phi = Y - Y\Phi^T \left(\Phi\Phi^T\right)^{-1} \Phi = Y \left(I - \Phi^T \left(\Phi\Phi^T\right)^{-1} \Phi\right) \tag{17.27}$$

and hence

$$\eta \left(Y - \hat{\Theta}\Phi\right) \left(Y - \hat{\Theta}\Phi\right)^T = \eta Y \left(I - \Phi^T \left(\Phi\Phi^T\right)^{-1} \Phi\right) Y^T \tag{17.28}$$

17.2 The Recursive Least-Squares Estimate

Suppose that we have already computed the least-squares estimate (17.3) of (17.9), when all of a sudden, somebody presents us with some more data $a \in \mathbb{C}^m$ and $b \in \mathbb{C}^p$ to consider. Specifically, suppose that we are now asked to compute the least-square estimate

$$\begin{aligned} \hat{\Theta}' &= \underset{\Theta}{\operatorname{argmin}} \left\| \Theta \left[\begin{array}{cc} A, & a \end{array} \right] - \left[\begin{array}{cc} B, & b \end{array} \right] \right\|_F && (17.29) \\ &= \left[\begin{array}{cc} B, & b \end{array} \right] \left[\begin{array}{c} A^T \\ a^T \end{array} \right] \left(\left[\begin{array}{cc} A, & a \end{array} \right] \left[\begin{array}{c} A^T \\ a^T \end{array} \right] \right)^{-1} && (17.30) \\ &= \left(BA^T + ba^T \right) \left(AA^T + aa^T \right)^{-1} && (17.31) \end{aligned}$$

Then fortunately, we can use the old estimate (17.3) to drastically simplify the computation. Specifically, using the matrix inversion lemma (Lemma B.1 in Appendix **??**), we find that the inverse $P' \triangleq \left(AA^T + aa^T \right)^{-1}$ is given by

$$P' \triangleq \left(AA^T + aa^T \right)^{-1} = \left(AA^T \right)^{-1} - \frac{\left(AA^T \right)^{-1} aa^T \left(AA^T \right)^{-1}}{1 + a^T \left(AA^T \right)^{-1} a} \qquad (17.32)$$

and hence the new estimate $\hat{\Theta}'$ is given by

$$\hat{\Theta}' = \left(BA^T + ba^T \right) \left[\left(AA^T \right)^{-1} - \frac{\left(AA^T \right)^{-1} aa^T \left(AA^T \right)^{-1}}{1 + a^T \left(AA^T \right)^{-1} a} \right] \qquad (17.33)$$

Although it may not be apparent, this result is quite powerful since we already computed $\left(AA^T \right)^{-1}$ when we calculated our initial estimate $\hat{\Theta}$ in (17.3). Hence instead of having to compute the $m \times m$ inverse $P' \triangleq \left(AA^T + aa^T \right)^{-1}$ from scratch, we are left with just a few matrix multiplications and a scalar division. To emphasize this point, let P denote the precalculated inverse $\left(AA^T \right)^{-1}$, that is, let

$$P \triangleq \left(AA^T \right)^{-1} \qquad (17.34)$$

Then $\hat{\Theta} = BA^T P$. Furthermore, from (17.32), we have that

$$P' \triangleq \left(AA^T + aa^T \right)^{-1} = P - \frac{Paa^T P}{1 + a^T Pa} \qquad (17.35)$$

and from (17.33), we find that

$$\hat{\Theta}' = \hat{\Theta}\left(I_m - \frac{aa^TP}{1+a^TPa}\right) + b\left(a^TP - \frac{a^TPaa^TP}{1+a^TPa}\right) \tag{17.36}$$

$$= \hat{\Theta}\left(I_m - \frac{aa^TP}{1+a^TPa}\right) + b\left(\frac{a^TP + a^TPa^TPa - a^TPaa^TP}{1+a^TPa}\right) \tag{17.37}$$

Finally, since a^TPa evaluates to a scalar, it commutes with a^TP, that is, $a^TPa^TPa = a^TPaa^TP$. Hence

$$\hat{\Theta}' = \hat{\Theta}\left(I_m - \frac{aa^TP}{1+a^TPa}\right) + b\left(\frac{a^TP + a^TPaa^TP - a^TPaa^TP}{1+a^TPa}\right) \tag{17.38}$$

$$= \hat{\Theta}\left(I_m - \frac{aa^TP}{1+a^TPa}\right) + b\left(\frac{a^TP}{1+a^TPa}\right) \tag{17.39}$$

Therefore, letting $k \triangleq \dfrac{a^TP}{1+a^TPa}$, it follows that

$$P' = P - Pak \tag{17.40}$$

$$\hat{\Theta}' = \hat{\Theta} + \left(b - \hat{\Theta}a\right)k \tag{17.41}$$

17.2.1 The Recursive Least-Squares Algorithm

The previous section showed how to update the least-squares estimate when one new piece of data was added. However, there is no reason why we can't apply this process sequentially, as more data is added. Specifically, suppose that we want to compute the estimate

$$\hat{\Theta}_N = \underset{\Theta}{\operatorname{argmin}} \left\| \Theta \left[\begin{array}{cccc} A & a_1 & \cdots & a_N \end{array}\right] - \left[\begin{array}{cccc} B & b_1 & \cdots & b_N \end{array}\right] \right\|_F \tag{17.42}$$

where the initial estimate

$$\hat{\Theta}_0 \triangleq \underset{\Theta}{\operatorname{argmin}} \left\| \Theta A - B \right\|_F = BA^T\left(AA^T\right)^{-1} \tag{17.43}$$

and the inverse $P_0 \triangleq \left(AA^T\right)^{-1}$ are known. Then letting $k_1 \triangleq \dfrac{a_1^T P_0}{1 + a_1^T P_0 a_1}$, from (17.41) we find that the estimate $\hat{\Theta}_1$ is given by

$$\hat{\Theta}_1 = \underset{\Theta}{\operatorname{argmin}} \left\| \Theta \left[\begin{array}{cc} A & a_1 \end{array} \right] - \left[\begin{array}{cc} B & b_1 \end{array} \right] \right\|_F \tag{17.44}$$

$$= \left(BA^T + b_1 a_1^T\right) \left(AA^T + a_1 a_1^T\right)^{-1} \tag{17.45}$$

$$= \hat{\Theta}_0 + \left(b_1 - \hat{\Theta}_0 a_1\right) k_1 \tag{17.46}$$

Furthemore, from (17.32), we find that the inverse $P_1 \triangleq \left(AA^T + a_1 a_1^T\right)^{-1}$ is given by

$$P_1 \triangleq \left(AA^T + a_1 a_1^T\right)^{-1} = P_0 - P_0 a_1 k_1 \tag{17.47}$$

Next, viewing P_1 and $\hat{\Theta}_1$ as the precomputed inverse and initial estimate, we can again apply the recursive least-squares update. Specifically, letting $k_2 \triangleq \dfrac{a_2^T P_1}{1 + a_2^T P_1 a_2}$, then from (17.41) we find that

$$\hat{\Theta}_2 = \underset{\Theta}{\operatorname{argmin}} \left\| \Theta \left[\begin{array}{ccc} A & a_1 & a_2 \end{array} \right] - \left[\begin{array}{ccc} B & b_1 & b_2 \end{array} \right] \right\|_F \tag{17.48}$$

$$= \left(\left(BA^T + b_1 a_1^T\right) + b_2 a_2^T\right)\left(\left(AA^T + a_1 a_1^T\right) + a_2 a_2^T\right)^{-1} \tag{17.49}$$

$$= \hat{\Theta}_1 + \left(b_2 - \hat{\Theta}_1 a_2\right) k_2 \tag{17.50}$$

where, from (17.32), we now have that

$$P_2 \triangleq \left(\left(AA^T + a_1 a_1^T\right) + a_2 a_2^T\right)^{-1} = P_1 - P_1 a_2 k_2 \tag{17.51}$$

Hence proceeding in this manner, we eventually obtain the estimate $\hat{\Theta}_N$. We summarize this procedure in the following algorithm:

```
Data: $A \in \mathbb{C}^{m\times n}$, $B \in \mathbb{C}^{p\times n}$, $a_1, \ldots, a_N \in \mathbb{C}^m$, $b_1, \ldots, b_N \in \mathbb{C}^p$
Result: The solution $\hat{\Theta}_N$ of the least-squares problem (17.42),
$\hat{\Theta}_N = \underset{\Theta}{\operatorname{argmin}} \left\| \Theta \begin{bmatrix} A & a_1 & \cdots & a_N \end{bmatrix} - \begin{bmatrix} B & b_1 & \cdots & b_N \end{bmatrix} \right\|_F$
Initialization:
    $P_0 = \left(AA^T\right)^{-1}$    (17.52)
    $\hat{\Theta}_0 = BA^T P_0$    (17.53)
Main:
    $i = 1$
    while ($i \leq N$) do
        $k_i = \frac{a_i^T P_{i-1}}{1 + a_i^T P_{i-1} a_i}$    (17.54)
        $P_i = P_{i-1} - P_{i-1} a_i k_i$    (17.55)
        $\hat{\Theta}_i = \hat{\Theta}_{i-1} + \left(b_i - \hat{\Theta}_{i-1} a_i\right) k_i$    (17.56)
        Increment $i$
    end
    return $\hat{\Theta}_N$
```

17.3 Recursive Least-Squares Initialization

If the recursive least-squares algorithm is initialized with P_0 and $\hat{\Theta}_0$ as in (17.52) and (17.53), then the result $\hat{\Theta}_N$ is the exact solution of the least-squares problem (17.42). However, there are two notable cases in which you might not want to initialize the algorithm in this way:

(1) If the dimension of AA^T is large, in which case, computing the inverse $P_0 = \left(AA^T\right)^{-1}$ is computationally expensive.

(2) If AA^T is not invertible, in which case, $P_0 = (AA^T)^{-1}$ does not exist.

To understand the consequences of not choosing the "correct" initial conditions for P_0 and $\hat{\Theta}_0$, we will develop P_i and $\hat{\Theta}_i$ as functions of arbitrary initial conditions P_0 and $\hat{\Theta}_0$.

First, from (17.32) and (17.55), note that if P_{i-1} is invertible, then

$$P_i = P_{i-1} - \frac{P_{i-1} a_i a_i^T P_{i-1}}{1 + a_i^T P_{i-1} a_i} = \left(P_{i-1}^{-1} + a_i a_i^T \right)^{-1} \tag{17.57}$$

from which we find that

$$\frac{a_i^T P_{i-1}}{1 + a_i^T P_{i-1} a_i} = a_i^T \left(P_{i-1}^{-1} + a_i a_i^T \right)^{-1} \tag{17.58}$$

$$\frac{a_i a_i^T P_{i-1}}{1 + a_i^T P_{i-1} a_i} = I - P_{i-1}^{-1} \left(P_{i-1}^{-1} + a_i a_i^T \right)^{-1} \tag{17.59}$$

Next, combining (17.56)-(17.59), we find that

$$\begin{aligned} \hat{\Theta}_i &= \hat{\Theta}_{i-1} + b_i \frac{a_i^T P_{i-1}}{1 + a_i^T P_{i-1} a_i} - \hat{\Theta}_{i-1} \frac{a_i a_i^T P_{i-1}}{1 + a_i^T P_{i-1} a_i} && (17.60) \\ &= \hat{\Theta}_{i-1} P_{i-1}^{-1} \left(P_{i-1}^{-1} + a_i a_i^T \right)^{-1} + b_i a_i^T \left(P_{i-1}^{-1} + a_i a_i^T \right)^{-1} && (17.61) \\ &= \left(\hat{\Theta}_{i-1} P_{i-1}^{-1} + b_i a_i^T \right) \left(P_{i-1}^{-1} + a_i a_i^T \right)^{-1} && (17.62) \\ &= \left(\hat{\Theta}_{i-1} P_{i-1}^{-1} + b_i a_i^T \right) P_i && (17.63) \end{aligned}$$

and therefore

$$\hat{\Theta}_i P_i^{-1} = \hat{\Theta}_{i-1} P_{i-1}^{-1} + b_i a_i^T \tag{17.64}$$

Furthermore, inverting (17.57), we have that

$$P_i^{-1} = P_{i-1}^{-1} + a_i a_i^T \tag{17.65}$$

Finally, from (17.64) and (17.65), note that

$$\hat{\Theta}_i P_i^{-1} = \hat{\Theta}_0 P_0^{-1} + \sum_{j=1}^{i} b_j a_j^T \tag{17.66}$$

$$P_i^{-1} = P_0^{-1} + \sum_{j=1}^{i} a_j a_j^T \tag{17.67}$$

Therefore, given the arbitrary initial conditions P_0 and $\hat{\Theta}_0$, the result $\hat{\Theta}_N$ of the recursive least-squares algorithm is given by

$$P_N = \left(P_0^{-1} + \sum_{j=1}^{N} a_j a_j^T \right)^{-1} \tag{17.68}$$

$$\hat{\Theta}_N = \left(\hat{\Theta}_0 P_0^{-1} + \sum_{j=1}^{N} b_j a_j^T \right) \left(P_0^{-1} + \sum_{j=1}^{N} a_j a_j^T \right)^{-1} \tag{17.69}$$

SECTION 18

Generalized Least-Squares

The least-squares problem (17.9) is sometimes called *ordinary least-squares* or *linear least-squares*. We can generalize it by allowing for the error $\Theta A - B$ to be weighted by an arbitrary matrix M, and by allowing the result to be biased towards an arbitrary value Θ_0 with weight L. Hence, a more general form of the least-squares problem is given by

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmin}} \left\| (\Theta A - B) M \right\|_F + \left\| (\Theta - \Theta_0) L \right\|_F \tag{18.1}$$

where M and L are rectangular matrices with consistent dimensions. Strangely enough, even though (18.1) is a generalized form of (17.9), it can be rearranged to look like the standard form (17.9) by noting that

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmin}} \left\| \Theta \begin{bmatrix} L, & AM \end{bmatrix} - \begin{bmatrix} \Theta_0 L, & BM \end{bmatrix} \right\|_F \tag{18.2}$$

which, as we already showed in (17.3), has the unique solution

$$\hat{\Theta} = \left(\Theta_0 L L^T + B M M^T A^T \right) \left(L L^T + A M M^T A^T \right)^{-1} \tag{18.3}$$

when $\begin{bmatrix} AM, & L \end{bmatrix}$ has full row rank.

18.1 Improperly Initialized Recursive Least-Squares

When the initial conditions P_0 and $\hat{\Theta}_0$ of the recursive least-squares algorithm are chosen as in (17.52) and (17.53), then the result $\hat{\Theta}_N$ is the exact solution (17.3) of the least-squares problem (17.9). But what about when the initial conditions P_0 and $\hat{\Theta}_0$ are chosen arbitrarily? Well it turns out that the recursive least-squares algorithm returns the solution of a generalized least-squares problem. Specifically, consider the generalized least-squares problem (18.1), where $M = I_i$, $A = \begin{bmatrix} \tilde{a}_1 & \cdots & \tilde{a}_i \end{bmatrix} \in \mathbb{C}^{m\times i}$, and $B = \begin{bmatrix} \tilde{b}_1 & \cdots & \tilde{b}_i \end{bmatrix} \in \mathbb{C}^{p\times i}$. Then the generalized least-squares estimate (18.3) is of the form

$$\hat{\Theta} = \left(\Theta_0 LL^T + \sum_{j=1}^{i} \tilde{b}_i \tilde{a}_i^T \right) \left(LL^T + \sum_{j=1}^{i} \tilde{b}_i \tilde{a}_i^T \right)^{-1} \tag{18.4}$$

Hence comparing (17.69) and (18.3), we see that the result $\hat{\Theta}_N$ of the recursive least-squares algorithm using the arbitrary initial conditions P_0 and $\hat{\Theta}_0$, is the solution of the generalized least-squares problem

$$\hat{\Theta} = \operatorname*{argmin}_{\Theta} \left\| \Theta \begin{bmatrix} \tilde{a}_1 \cdots \tilde{a}_N \end{bmatrix} - \begin{bmatrix} \tilde{b}_1 \cdots \tilde{b}_N \end{bmatrix} \right\|_F + \left\| (\Theta - \hat{\Theta}_0) L \right\|_F \tag{18.5}$$

$$= \operatorname*{argmin}_{\Theta} \left\| \Theta \begin{bmatrix} L & \tilde{a}_1 & \cdots & \tilde{a}_N \end{bmatrix} - \begin{bmatrix} \hat{\Theta}_0 L & \tilde{b}_1 & \cdots & \tilde{b}_N \end{bmatrix} \right\|_F \tag{18.6}$$

where $LL^T = P_0^{-1}$, that is, L denotes the Cholesky decomposition of P_0^{-1}.

18.2 Generalized Recursive Least-Squares

The previous section showed that the recursive least-squares algorithm returns the solution of the generalized least-squares problem (18.5), that is, when $M = I$ in (18.1). Next, we show how to solve the more general least-squares problem (18.1) recursively when $M \neq I$.

Suppose that we have computed the generalized least-squares solution

$$\hat{\Theta}_0 = \operatorname*{argmin}_{\Theta} \left\| (\Theta A - B) M \right\|_F + \left\| (\Theta - \Theta_0) L \right\|_F \tag{18.7}$$

$$= \left(\Theta_0 LL^T + BMM^T A^T \right) \left(LL^T + AMM^T A^T \right)^{-1} \tag{18.8}$$

where the inverse $P_0 \triangleq \left(LL^T + AMM^T A^T \right)^{-1}$ is known. Furthermore, suppose that we are given the addition data $a_1 \in \mathbb{C}^m$ and $b_1 \in \mathbb{C}^p$, and asked to solve the new generalized least-squares problem

$$\hat{\Theta}_1 = \operatorname*{argmin}_{\Theta} \left\| \left(\Theta \left[A, \;\; a_1 \right] - \left[B, \;\; b_1 \right] \right) \begin{bmatrix} M & 0 \\ 0 & m_1 \end{bmatrix} \right\|_F + \left\| (\Theta - \Theta_0) L \right\|_F$$

that is, suppose that we are asked to compute

$$\begin{aligned}\hat{\Theta}_1 &= \left(\Theta_0 LL^T + BMM^T A^T + m_1^2 b_1 a_1^T\right)\left(LL^T + AMM^T A^T + m_1^2 b_1 a_1^T\right)^{-1} \\ &= \left(B_0 A_0^T + m_1^2 b_1 a_1^T\right)\left(A_0 A_0^T + m_1^2 a_1 a_1^T\right)^{-1} \qquad (18.9)\end{aligned}$$

where $A_0 \triangleq \left[\begin{array}{cc} L, & AM \end{array}\right]$ and $B_0 \triangleq \left[\begin{array}{cc} \Theta_0 L, & BM \end{array}\right]$. Then from (17.32), we find that

$$\left(A_0 A_0^T + m_1^2 a_1 a_1^T\right)^{-1} = \left(A_0 A_0^T\right)^{-1} - \frac{\left(A_0 A_0^T\right)^{-1} a_1 a_1^T \left(A_0 A_0^T\right)^{-1}}{1/m_1^2 + a_1^T \left(A_0 A_0^T\right)^{-1} a_1} \qquad (18.10)$$

Therefore, like in the standard recursive least-squares case, we find that

$$k_1 = \frac{a_1^T P_0}{1/m_1^2 + a_1^T P_0 a_1} \qquad (18.11)$$

$$P_1 = \left(A_0 A_0^T + m_1^2 a_1 a_1^T\right)^{-1} = P_0 - P_0 a_1 k_1 \qquad (18.12)$$

$$\hat{\Theta}_1 = \left(B_0 A_0^T + m_1^2 b_1 a_1^T\right) P_1 = \hat{\Theta}_0 + \left(b_1 - \hat{\Theta}_0 a_1\right) k_1 \qquad (18.13)$$

Hence the generalized least-squares problem

$$\hat{\Theta}_N = \underset{\Theta}{\operatorname{argmin}} \left\| \left(\Theta \tilde{A} - \tilde{B}\right) \tilde{M} \right\|_F + \left\| (\Theta - \Theta_0) L \right\|_F \qquad (18.14)$$

$$\tilde{A} \triangleq \left[\begin{array}{cccc} A, & a_1, & \cdots, & a_N \end{array}\right] \qquad (18.15)$$

$$\tilde{B} \triangleq \left[\begin{array}{cccc} B, & b_1, & \cdots, & b_N \end{array}\right] \qquad (18.16)$$

$$\tilde{M} \triangleq \begin{bmatrix} M & & & \\ & m_1 & & \\ & & \ddots & \\ & & & m_N \end{bmatrix} \qquad (18.17)$$

can be solved by the modified recursive least-squares algorithm:

```
Data: $\tilde{A}$, $\tilde{B}$, $\tilde{M}$, $\Theta_0$, and $L$ as defined in (18.14)-(18.17).
Result: The solution $\hat{\Theta}_N$ of the generalized least-squares
  problem (18.14), that is,
        $\hat{\Theta}_N = \underset{\Theta}{\operatorname{argmin}} \left\| \left(\Theta \tilde{A} - \tilde{B}\right) \tilde{M} \right\|_F + \left\| \left(\Theta - \Theta_0\right) L \right\|_F$   (18.18)
Initialization:
        $P_0 = \left(LL^T + AMM^TA^T\right)^{-1}$   (18.19)
        $\hat{\Theta}_0 = \left(\Theta_0 LL^T + BMM^TA^T\right) P_0$   (18.20)
 Main:
   $i = 1$
   while $(i \leq N)$ do
        $k_i = \frac{a_i^T P_{i-1}}{1/m_i^2 + a_i^T P_{i-1} a_i}$   (18.21)
        $P_i = P_{i-1} - P_{i-1} a_i k_i$   (18.22)
        $\hat{\Theta}_i = \hat{\Theta}_{i-1} + \left(b_i - \hat{\Theta}_{i-1} a_i\right) k_i$   (18.23)
        Increment $i$
   end
   return $\hat{\Theta}_N$
```

SECTION 19

Forgetting Factors

Weighting the data with a matrix $\tilde{M}$, like we did in the generalized least-squares problem (18.18), allows us to emphasize the importance of certain pieces of data. For instance, it will typically happen that the most current data a_N and b_N are also the most important, in which case, we want to choose a high value of m_N to emphasize their importance relative to the other pieces of data. However, this quickly leads to a problem. Specifically, given more and more data, where the newest data is continually the most important, we have to increase the weights m without bound so that the newest data is continually emphasized the most, that is,

$$\lim_{N\to\infty} m_N \to \infty \tag{19.1}$$

A better way forward is to introduce what are called *forgetting factors.* The idea is simply that instead of emphasizing the new data by choosing a larger and larger value for m_N, we can deemphasize the old data, that is, we can slowly *forget* the old data. We can accomplish this by choosing a constant $\lambda < 1$, called a *forgetting factor*, and to instead consider the least-squares problems

$$\hat{\Theta}_0 = \underset{\Theta}{\operatorname{argmin}} \left\| \Theta \left[\begin{array}{c} A \end{array} \right] - \left[\begin{array}{c} B \end{array} \right] \right\|_F \tag{19.2}$$

$$\hat{\Theta}_1 = \underset{\Theta}{\operatorname{argmin}} \left\| \Theta \left[\begin{array}{cc} \lambda A, & a_1 \end{array} \right] - \left[\begin{array}{cc} \lambda B, & b_1 \end{array} \right] \right\|_F \tag{19.3}$$

$$\hat{\Theta}_2 = \underset{\Theta}{\operatorname{argmin}} \left\| \Theta \left[\begin{array}{ccc} \lambda^2 A, & \lambda a_1, & a_2 \end{array} \right] - \left[\begin{array}{ccc} \lambda^2 B, & \lambda b_1, & b_2 \end{array} \right] \right\|_F \tag{19.4}$$

that is, the general least-squares problem with a forgetting factor $\lambda < 1$ is given by

$$\hat{\Theta}_N = \underset{\Theta}{\operatorname{argmin}} \left\| \Theta A_N - B_N \right\|_F \tag{19.5}$$

$$A_N \triangleq \left[\begin{array}{cccc} \lambda^N A, & \lambda^{N-1} a_1, & \cdots, & a_N \end{array} \right] \tag{19.6}$$

$$B_N \triangleq \left[\begin{array}{cccc} \lambda^N B, & \lambda^{N-1} b_1, & \cdots, & b_N \end{array} \right] \tag{19.7}$$

Hence unlike in the generalized least-squares algorithm, where the weighting matrix M was fixed for all time, the forgetting factor approach uses a variable weighting matrix M, which gradually deemphasizes the old data.

Fortunately, the forgetting factor approach (19.5) also permits a simple recursive solution. Specifically, suppose that the inverse $P_0 \triangleq \left(AA^T\right)^{-1}$ and the initial solution

$$\hat{\Theta}_0 = \underset{\Theta}{\operatorname{argmin}} \left\| \Theta A - B \right\|_F = BA^T \left(AA^T\right)^{-1} = BA^T P_0 \tag{19.8}$$

are known. Then from (17.32), we find that

$$
\begin{aligned}
P_1 &\triangleq \left(\lambda^2 AA^T + a_1 a_1^T\right)^{-1} && (19.9)\\
&= \left(\lambda^2 AA^T\right)^{-1} - \frac{\left(\lambda^2 AA^T\right)^{-1} a_1 a_1^T \left(\lambda^2 AA^T\right)^{-1}}{1 + a_1^T \left(\lambda^2 AA^T\right)^{-1} a_1} && (19.10)\\
&= \frac{1}{\lambda^2}\left(P_0 - \frac{P_0 a_1 a_1^T P_0}{\lambda^2 + a_1^T P_0 a_1}\right) && (19.11)
\end{aligned}
$$

Hence the solution $\hat{\Theta}_1$ of (19.3) is given by

$$
\begin{aligned}
\hat{\Theta}_1 &= \left(\lambda^2 BA^T + b_1 a_1^T\right)\left(\lambda^2 AA^T + a_1 a_1^T\right)^{-1} && (19.12)\\
&= \left(BA^T + \frac{1}{\lambda^2} b_1 a_1^T\right)\left(P_0 - \frac{P_0 a_1 a_1^T P_0}{\lambda^2 + a_1^T P_0 a_1}\right) && (19.13)\\
&= \hat{\Theta}_0 \left(I - \frac{a_1 a_1^T P_0}{\lambda^2 + a_1^T P_0 a_1}\right) + \frac{1}{\lambda^2} b_1 \left(a_1^T P_0 - \frac{a_1^T P_0 a_1 a_1^T P_0}{\lambda^2 + a_1^T P_0 a_1}\right) && (19.14)\\
&= \hat{\Theta}_0 \left(I - \frac{a_1 a_1^T P_0}{\lambda^2 + a_1^T P_0 a_1}\right) + \frac{1}{\lambda^2} b_1 \left(\frac{\lambda^2 a_1^T P_0}{\lambda^2 + a_1^T P_0 a_1}\right) && (19.15)
\end{aligned}
$$

Therefore, letting $k_1 \triangleq \dfrac{a_1^T P_0}{\lambda^2 + a_1^T P_0 a_1}$, we find that

$$
\begin{aligned}
P_1 &\triangleq \left(\lambda^2 AA^T + a_1 a_1^T\right)^{-1} = \frac{1}{\lambda^2}\left(P_0 - P_0 a_1 k_1\right) && (19.16)\\
\hat{\Theta}_1 &\triangleq \left(\lambda^2 BA^T + b_1 a_1^T\right) P_1 = \hat{\Theta}_0 + \left(b_1 - \hat{\Theta}_0 a_1\right) k_1 && (19.17)
\end{aligned}
$$

```
Data: $A \in \mathbb{C}^{m \times n}$, $B \in \mathbb{C}^{p \times n}$, $a_1, \dots, a_N \in \mathbb{C}^m$, $b_1, \dots, b_N \in \mathbb{C}^p$
Result: The solution $\hat{\Theta}_N$ of the least-squares problem
    $\hat{\Theta}_N = \operatorname*{argmin}_{\Theta} \left\| \Theta A_N - B_N \right\|_F$    (19.18)
    $A_N \triangleq \begin{bmatrix} \lambda^N A, & \lambda^{N-1} a_1, & \cdots, & a_N \end{bmatrix}$    (19.19)
    $B_N \triangleq \begin{bmatrix} \lambda^N B, & \lambda^{N-1} b_1, & \cdots, & b_N \end{bmatrix}$    (19.20)
  where $\lambda < 1$ is called the *forgetting factor*.
Initialization:
    $P_0 = \left( A A^T \right)^{-1}$    (19.21)
    $\hat{\Theta}_0 = B A^T P_0$    (19.22)
 Main:
  $i = 1$
  while $(i \leq N)$ do
      $k_i = \dfrac{a_i^T P_{i-1}}{\lambda^2 + a_i^T P_{i-1} a_i}$    (19.23)
      $P_i = \dfrac{1}{\lambda^2} \left( P_{i-1} - P_{i-1} a_i k_i \right)$    (19.24)
      $\hat{\Theta}_i = \hat{\Theta}_{i-1} + \left( b_i - \hat{\Theta}_{i-1} a_i \right) k_i$    (19.25)
      Increment $i$
  end
  return $\hat{\Theta}_N$
```

SECTION 20

The Partitioned Least-Squares Estimate

Consider the least-squares estimate

$$\begin{bmatrix} \hat{\alpha} & \hat{\beta} \end{bmatrix} = \operatorname*{argmin}_{\alpha, \beta} \left\| \begin{bmatrix} \alpha & \beta \end{bmatrix} \begin{bmatrix} A \\ B \end{bmatrix} - C \right\|_2 \tag{20.1}$$

which is the solution of

$$\begin{bmatrix} \alpha & \beta \end{bmatrix} \begin{bmatrix} A \\ B \end{bmatrix} \begin{bmatrix} A^T & B^T \end{bmatrix} = C \begin{bmatrix} A^T & B^T \end{bmatrix} \tag{20.2}$$

that is,

$$\hat{\alpha} AA^T + \hat{\beta} BA^T = CA^T \tag{20.3}$$

$$\hat{\alpha} AB^T + \hat{\beta} BB^T = CB^T \tag{20.4}$$

Hence if AA^T and BB^T are invertible, then

$$\hat{\alpha} = \left(CA^T - \hat{\beta} BA^T \right) \left(AA^T \right)^{-1} \tag{20.5}$$

$$\hat{\beta} = \left(CB^T - \hat{\alpha} AB^T \right) \left(BB^T \right)^{-1} \tag{20.6}$$

and therefore

$$\hat{\alpha} A \left[I - B^T \left(BB^T \right)^{-1} B \right] A^T = C \left[I - B^T \left(BB^T \right)^{-1} B \right] A^T \tag{20.7}$$

$$\hat{\beta} B \left[I - A^T \left(AA^T \right)^{-1} A \right] B^T = C \left[I - A^T \left(AA^T \right)^{-1} A \right] B^T \tag{20.8}$$

SECTION 21

Least-Squares in Regression

21.1 Linear Least-Squares Scenario

Consider the linear model

$$\begin{aligned} y(k+n) + \alpha_{n-1} y(k+n-1) + \cdots + \alpha_0 y(k) \\ = \beta_n u(k+n) + \beta_{n-1} u(k+n-1) + \cdots + \beta_0 u(k) \end{aligned} \tag{21.1}$$

where $\alpha_0, \ldots, \alpha_{n-1}$ and $\beta_0, \ldots, \beta_n$ are unknown. Then letting

$$\theta \triangleq \begin{bmatrix} \alpha_{n-1} & \cdots & \alpha_0 & \beta_n & \cdots & \beta_0 \end{bmatrix} \tag{21.2}$$

$$\phi_n(k) \triangleq \begin{bmatrix} -y(k+n-1) \\ \vdots \\ -y(k) \\ u(k+n) \\ \vdots \\ u(k) \end{bmatrix} \tag{21.3}$$

we find that

$$y(k+n) = \theta\phi_n(k) \tag{21.4}$$

Furthermore, if we sample the signals u and y at the points $k = 1, \ldots, N$, then we find that

$$\left[\begin{array}{ccc} y(n+1) & \cdots & y(N) \end{array}\right] = \theta \left[\begin{array}{ccc} \phi_n(1) & \cdots & \phi_n(N-n) \end{array}\right] \tag{21.5}$$

Why can't we use $\phi(N)$? We simply don't have the data.

Hence letting

$$\begin{aligned} Y_{n,N} &\triangleq \left[\begin{array}{ccc} y(n+1) & \cdots & y(N) \end{array}\right] \qquad (21.6)\\ \Phi_{n,N} &\triangleq \left[\begin{array}{ccc} \phi_n(1) & \cdots & \phi_n(N{-}n) \end{array}\right] \qquad (21.7) \end{aligned}$$

we have that

$$Y_{n,N} = \theta\Phi_{n,N} \tag{21.8}$$

in which case, θ can be found by

$$\theta = Y_{n,N}\Phi_{n,N}^T\left(\Phi_{n,N}\Phi_{n,N}^T\right)^{-1} \tag{21.9}$$

Next, let's assume that there is noise in the model, that is,

$$\begin{aligned} & y(k+n) + \alpha_{n-1}y(k+n-1) + \cdots + \alpha_0 y(k) \\ &= \beta_n u(k+n) + \beta_{n-1}u(k+n-1) + \cdots + \beta_0 u(k) \\ &+ \gamma_n w(k+n) + \gamma_{n-1}w(k+n-1) + \cdots + \gamma_0 w(k) \end{aligned} \tag{21.10}$$

Then letting

$$\begin{aligned} \tilde{\phi}_n(k) &\triangleq \left[\begin{array}{c} w(k+n) \\ \vdots \\ w(k) \end{array}\right] \qquad (21.11)\\ \tilde{\theta} &\triangleq \left[\begin{array}{ccc} \gamma_n & \cdots & \gamma_0 \end{array}\right] \qquad (21.12)\\ \tilde{\Phi}_{n,N} &\triangleq \left[\begin{array}{ccc} \tilde{\phi}_n(1) & \cdots & \tilde{\phi}_n(N-n) \end{array}\right] \qquad (21.13) \end{aligned}$$

we have that

$$Y_{n,N} = \theta\Phi_{n,N} + \tilde{\theta}\tilde{\Phi}_{n,N} \tag{21.14}$$

where θ, $\tilde{\theta}$, and $\tilde{\Phi}_{n,N}$ are all unknown. Hence letting

$$\hat{\theta} = Y_{n,N}\Phi_{n,N}^T \left(\Phi_{n,N}\Phi_{n,N}^T\right)^{-1} \tag{21.15}$$

we find that

$$\begin{align} \hat{\theta} &= \left(\theta\Phi_{n,N} + \tilde{\theta}\tilde{\Phi}_{n,N}\right)\Phi_{n,N}^T \left(\Phi_{n,N}\Phi_{n,N}^T\right)^{-1} \tag{21.16}\\ &= \theta + \tilde{\theta}\tilde{\Phi}_{n,N}\Phi_{n,N}^T \left(\Phi_{n,N}\Phi_{n,N}^T\right)^{-1} \tag{21.17}\end{align}$$

Example 21.1 Consider the model

$$y(k) = \beta_0 u(k) + \gamma_0 w(k) \tag{21.18}$$

where $u(k)$ is the sinusoidal signal

$$u(k) = \sin(2\pi f k + \phi) \tag{21.19}$$

and $w(k)$ is the white Gaussian noise process. Then the least-squares estimate $\hat{\beta}_0$ of β_0 is given by

$$\hat{\beta}_0 = \left(\sum_{k=1}^{N} y(k)u(k)\right) \bigg/ \left(\sum_{k=1}^{N} u^2(k)\right) \tag{21.20}$$

which,

$$\begin{align} \hat{\beta}_0 &= \left(\sum_{k=1}^{N} \Big[\beta_0 u(k) + \gamma_0 w(k)\Big] u(k)\right) \bigg/ \left(\sum_{k=1}^{N} u^2(k)\right) \tag{21.21}\\ &= \left(\left[\sum_{k=1}^{N} \beta_0 u^2(k)\right] + \left[\sum_{k=1}^{N} \gamma_0 w(k)u(k)\right]\right) \bigg/ \left(\sum_{k=1}^{N} u^2(k)\right) \tag{21.22}\\ &= \beta_0 + \gamma_0 \left(\sum_{k=1}^{N} w(k)u(k)\right) \bigg/ \left(\sum_{k=1}^{N} u^2(k)\right) \tag{21.23}\end{align}$$

Hence

$$\begin{align} \mathbb{E}\big[\hat{\beta}_0\big] &= \beta_0 + \gamma_0 \mathbb{E}\left[\sum_{k=1}^{N} w(k)u(k)\right] \bigg/ \left(\sum_{k=1}^{N} u^2(k)\right) \tag{21.24}\\ &= \beta_0 + \gamma_0 \left(\sum_{k=1}^{N} \mathbb{E}\Big[w(k)\Big] u(k)\right) \bigg/ \left(\sum_{k=1}^{N} u^2(k)\right) \tag{21.25}\end{align}$$

Therefore, since

$$\mathbb{E}\big[\hat{\beta}_0\big] = \beta_0 \tag{21.26}$$

But what about the variance, that is,

$$\mathbb{E}\Big[\Big(\hat{\beta}_0 - \mathbb{E}\big[\hat{\beta}_0\big]\Big)^2\Big] = \mathbb{E}\left[\gamma_0^2\left(\sum_{k=1}^{N} w(k)u(k)\right)^2 \Big/ \left(\sum_{k=1}^{N} u^2(k)\right)^2\right] \tag{21.27}$$

$$= \gamma_0^2\,\mathbb{E}\left[\left(\sum_{k=1}^{N} w(k)u(k)\right)^2\right] \Big/ \left(\sum_{k=1}^{N} u^2(k)\right)^2 \tag{21.28}$$

where, since the random process $\mathcal{W}$ is independent and identically distributed, then for all nonzero $i \in \mathbb{Z}$,

$$\mathbb{E}\Big[w(k)w(k+i)\Big] = 0 \tag{21.29}$$

Therefore

$$\mathbb{E}\Big[w(k)u(k)w(k+i)u(k+i)\Big] = \mathbb{E}\Big[w(k)w(k+i)\Big]u(k)u(k+i) = 0 \tag{21.30}$$

and hence

$$\mathbb{E}\Big[\Big(\hat{\beta}_0 - \mathbb{E}\big[\hat{\beta}_0\big]\Big)^2\Big] = \gamma_0^2\left(\sum_{k=1}^{N}\mathbb{E}\left[w^2(k)\right]u^2(k)\right) \Big/ \left(\sum_{k=1}^{N} u^2(k)\right)^2 \tag{21.31}$$

Finally, since $\mathcal{W}$ is stationary, then letting σ_w^2 denote the variance of $\mathcal{W}$, we find that

$$\mathbb{E}\Big[\Big(\hat{\beta}_0 - \mathbb{E}\big[\hat{\beta}_0\big]\Big)^2\Big] = \frac{\gamma_0^2\sigma_w^2}{\sum\limits_{k=1}^{N} u^2(k)} \leq \frac{2\gamma_0^2\sigma_w^2}{N - \Big|\csc\big(2\pi f\big)\Big|} \tag{21.32}$$

Hence the variance of $\hat{\beta}_0$ decreases as N increases, provided that $u(k) \neq 0$, in which case $\sin(2\pi f) = 0$. ◻

21.2 General Properties

Now, since w is a realization of a random process, $\hat{\theta}$ is a random variable. Hence

$$\mathbb{E}\big[\hat{\theta}\big] = \theta + \tilde{\theta}\mathbb{E}\left[\tilde{\Phi}_{n,N}\Phi_{n,N}^T\left(\Phi_{n,N}\Phi_{n,N}^T\right)^{-1}\right] \tag{21.33}$$

Unfortunately, the term inside the interval is really hard to analyze, so instead we analyze the matrices $\tilde{\Phi}_{n,N}\Phi_{n,N}^T$ and $\Phi_{n,N}\Phi_{n,N}^T$ separately, which are of the form

$$\tilde{\Phi}_{n,N}\Phi_{n,N}^T = \begin{bmatrix} \tilde{\phi}_n(1) & \cdots & \tilde{\phi}_n(N-n) \end{bmatrix}\begin{bmatrix} \phi_n^T(1) \\ \vdots \\ \phi_n^T(N-n) \end{bmatrix} = \sum_{k=1}^{N-n} \tilde{\phi}_n(k)\phi_n^T(k)$$

$$\Phi_{n,N}\Phi_{n,N}^T = \begin{bmatrix} \phi_n(1) & \cdots & \phi_n(N-n) \end{bmatrix}\begin{bmatrix} \phi_n^T(1) \\ \vdots \\ \phi_n^T(N-n) \end{bmatrix} = \sum_{k=1}^{N-n} \phi_n(k)\phi_n^T(k)$$

where

$$\phi_n(k)\phi_n^T(k) = \begin{bmatrix} y(k+n-1)y(k+n-1)^T & \cdots & -y(k+n-1)u(k)^T \\ \vdots & & \vdots \\ -u(k)y(k+n-1)^T & \cdots & u(k)u(k)^T \end{bmatrix}$$

$$\tilde{\phi}_n(k)\phi_n^T(k) = \begin{bmatrix} -w(k+n)y(k+n-1)^T & \cdots & w(k+n)u(k)^T \\ \vdots & & \vdots \\ -w(k)y(k+n-1)^T & \cdots & w(k)u(k)^T \end{bmatrix}$$

Specifically, we want to analyze the limiting behavior of terms of the form

$$\frac{1}{N}\sum_{k=1}^{N} y(k+i)y^T(k), \quad \frac{1}{N}\sum_{k=1}^{N} y(k+i)u^T(k)$$

$$\frac{1}{N}\sum_{k=1}^{N} w(k+i)y^T(k), \quad \frac{1}{N}\sum_{k=1}^{N} w(k+i)u^T(k)$$

Example 21.2 Consider the model

$$y(k+1) + \alpha_0 y(k) = \beta_1 u(k+1) + \beta_0 u(k) + \gamma_1 w(k+1) + \gamma_0 w(k) \tag{21.34}$$

where $u(k)$ is the sinusoidal signal

$$u(k) = \sin(2\pi f k + \phi) \tag{21.35}$$

and $w(k)$ is the white Gaussian noise process. Then the unknown parameter vector is given by

$$\theta \triangleq \begin{bmatrix} \alpha_0 & \beta_1 & \beta_0 \end{bmatrix} \tag{21.36}$$

where the least-squares estimate $\hat{\theta}$ of θ is given by

$$\begin{aligned}
\hat{\theta} &= \theta + \tilde{\theta}\tilde{\Phi}_{n,N}\Phi_{n,N}^T \left(\Phi_{n,N}\Phi_{n,N}^T\right)^{-1} \\
\tilde{\Phi}_{n,N}\Phi_{n,N}^T &= \sum_{k=1}^{N-1} \begin{bmatrix} w(k+1)y(k) & w(k+1)u(k+1) & w(k+1)u(k) \\ w(k)y(k) & w(k)u(k+1) & w(k)u(k) \end{bmatrix} \\
\Phi_{n,N}\Phi_{n,N}^T &= \sum_{k=1}^{N-1} \begin{bmatrix} y(k)y(k) & y(k)u(k+1) & y(k)u(k) \\ u(k+1)y(k) & u(k+1)u(k+1) & u(k+1)u(k) \\ u(k)y(k) & u(k)u(k+1) & u(k)u(k) \end{bmatrix}
\end{aligned}$$

In the previous example, we could easily determine that the least-squares estimate was unbiased, since we had that

$$\mathbb{E}\left[\tilde{\Phi}_{n,N}\Phi_{n,N}^T \left(\Phi_{n,N}\Phi_{n,N}^T\right)^{-1}\right] = \mathbb{E}\left[\tilde{\Phi}_{n,N}\Phi_{n,N}^T\right] \left(\Phi_{n,N}\Phi_{n,N}^T\right)^{-1} \tag{21.37}$$

This was possible since $\Phi_{n,N}$ was only a function of the deterministic input u. However, now we have that the $\Phi_{n,N}$ is also a function of the output, y, which is dependent on the random variables $\mathcal{W}$. Hence y is also a random variable. Furthermore, note that $\tilde{\Phi}_{n,N}\Phi_{n,N}^T$ and $\Phi_{n,N}\Phi_{n,N}^T$ are generally not independent. Hence we also have that, in general,

$$\mathbb{E}\left[\tilde{\Phi}_{n,N}\Phi_{n,N}^T \left(\Phi_{n,N}\Phi_{n,N}^T\right)^{-1}\right] \neq \mathbb{E}\left[\tilde{\Phi}_{n,N}\Phi_{n,N}^T\right] \mathbb{E}\left[\left(\Phi_{n,N}\Phi_{n,N}^T\right)^{-1}\right] \tag{21.38}$$

So where do we go from here? Basically, if you really want to know the properties of your estimate, then you can either

□

PART V

Models with Noise

SECTION 22

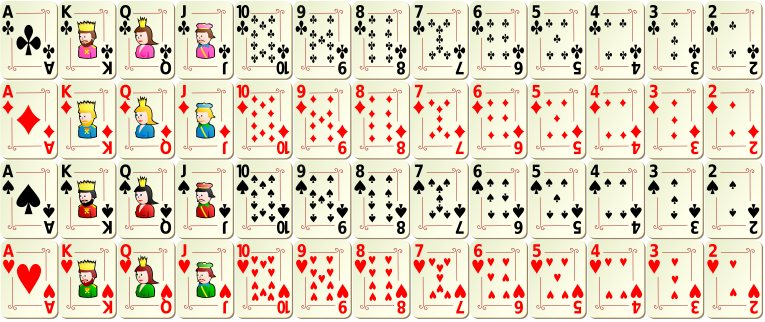
Probability

The purpose of probability theory is to quantify the likelihood of an event occurring as the result of an experiment. Specifically, we denote the probability of an event A occurring by $P(A)$, where

$$0 \leq P(A) \leq 1 \tag{22.1}$$

where the more likely A is to happen, the closer $P(A)$ is to 1.

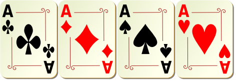
Example 22.1 Suppose we are drawing a single card from a well-shuffled deck of 52 cards.



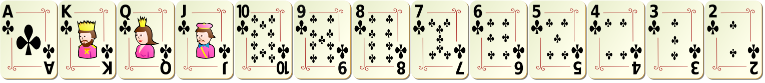
Then we might want to know the probabilities of the following events:

(A) The card will be an ace of clubs.

(B) The card will be an ace.



(C) The card will be in clubs.



Fortunately, if all of cards are equally likely to be drawn, then this is a relatively simple example. Specifically, you might already know that the probabilities of A, B, and C can be calculated by dividing the number of cards in the event by the total number of cards in the deck, that is,

$$P(A) = \frac{1}{52}, \qquad P(B) = \frac{4}{52} = \frac{1}{13}, \qquad P(C) = \frac{13}{52} = \frac{1}{4} \tag{22.2}$$

where the fact that $P(C) > P(A)$ tells us that C is more likely than A, that is, we are more likely to draw a club than we are to specifically draw the ace of clubs. ❒

22.1 Outcomes and Events

Although probability theory concerns itself with quantifying the likelihood of events, events are themselves composed of even smaller units, called *outcomes*, where:

Definition 22.1

- Every *experiment* has exactly one result.
- An *outcome* is a possible result of an experiment.
- An *event* is a set of outcomes.[3]
- The *sample space* Ω is the set of all of the possible outcomes of an experiment.

❒

When it is possible, we will denote the outcomes of an experiment by $\omega_1, \omega_2, \ldots$, where since Ω denotes the set of all possible outcomes, we have that

$$\Omega = \{\omega_1, \omega_2, \ldots\} \tag{22.3}$$

Example 22.2 In our card-choosing example, Example 22.1, there are 52 possible outcomes of drawing a single card from a deck of 52 cards. Hence the sample space contains 52 outcomes, that is, $\Omega = \{\omega_1, \ldots, \omega_{52}\}$. Furthermore, the events A, B, and C are sets of 1, 4, and 13 outcomes, respectively. ❒

[3] By *set* we mean an unordered collection of unique objects, that is, no two elements of the collection are the same.

Example 22.3 Suppose we are going to roll a standard 6-sided die. Then the sample space contains 6 possible outcomes, that is,

$\Omega = \{\omega_1, \ldots, \omega_6\} = \Big\{$ ⚀ , ⚁ , ⚂ , ⚃ , ⚄ , ⚅ $\Big\}$ (22.4)

Furthermore, depending on what game we are playing, we could be interested in the following events:

A = { ⚅ }: the roll will be a six

B = { ⚁ , ⚃ , ⚅ }: the roll will be even

C = { ⚂ , ⚃ , ⚄ , ⚅ }: the roll will be greater than two

which contain 1, 3, and 4 outcomes, respectively.

22.2 The Probability of an Event

Since we are defining events as sets of outcomes, let's look at how we should interpret the probability of an event. Specifically, let x denote the result of an experiment, and let $E \triangleq \{\omega_1, \ldots, \omega_n\}$ denote an event. Then the probability of E is the probability that the result is equal to one of the outcomes in E, that is,

$$P(E) = P\Big(x = \omega_1, \quad \text{or} \quad x = \omega_2, \quad \ldots, \quad \text{or} \quad x = \omega_n\Big) \tag{22.5}$$

Furthermore, since every experiment has exactly one distinct outcome, no more than one of the outcomes in E will occur. For instance, ω_1 and ω_2 cannot both occur.

Example 22.4 Let x denote a card that is drawn from a standard deck of 52 cards. Then the event that the card will be an ace is written as B = { 🃑 , 🃁 , 🂡 , 🂱 }, and the probability of B is the probability that the result is equal to one of the 4 outcomes in B, that is,

$P(B) = P\Big(x =$ 🃑 , or $x =$ 🃁 , or $x =$ 🂡 , or $x =$ 🂱 $\Big)$

Example 22.5 Let x denote the result of rolling a six-sided die. Then the event that the result is even is written as $B = \left\{ \text{⚁}, \text{⚃}, \text{⚅} \right\}$, and the probability of B is the probability that the result is equal to one of the 3 outcomes in B, that is,

$$P(B) = P\left(x = \text{⚁}\ ,\quad \text{or} \quad x = \text{⚃}\ ,\quad \text{or} \quad x = \text{⚅}\right)$$

□

22.3 First Guess at the Probability Measure

Since events are built from outcomes, and only one outcome can occur as the result of an experiment, it seems like we should be able to compute the probability of an event $E \triangleq \{\omega_1, \ldots, \omega_n\}$ from the probabilities $P(\omega_1), \ldots, P(\omega_n)$. The problem is that, aside from Example 22.1, all that we have said so far is that the range of P lies between 0 and 1. Hence the definition of a probability measure[4] P is actually quite ambiguous at the moment. However, before we get to the common definition of a probability measure (which we introduce later in Definition 22.9), we are going to examine the consequences of the following tentative definition, which is a simplified definition that will suffice for many cases.[5,6]

Tentative Definition 22.1 P is a *probability measure* if

(i) $P(\Omega) = 1$

and (ii) for every subset $\{\omega_1, \ldots, \omega_n\}$ of Ω,

$$P\Big(\{\omega_1, \ldots, \omega_n\}\Big) = P(\omega_1) + \cdots + P(\omega_n) \geq 0 \tag{22.6}$$

□

[4] Pretend that I wrote the word *function* if you do not know what a *measure* is.

[5] Tentative Definition 22.1 is not something that can be inferred from (22.1), or from the definition of an outcome or an event; it is a rule that we are choosing for the *probability measure* P. Furthermore, we call this a *tentative definition* since although this definition is sufficient for many circumstances, it is not sufficient for all circumstances. Hence we will need to change the definition later (see Definition 22.9).

[6] The specific values of $P(\omega_1), P(\omega_2), \ldots$ are not specified by Tenative Definition 22.1 since those values will depend on the specific experiment. However, if you had to assign the probabilities $P(\omega_1), P(\omega_2), \ldots$ to specific outcomes of an experiment, then you should check to make sure that your choices satisfy Tentative Definition 22.1.

To understand the statement $P(\Omega) = 1$, note that Ω is a set of outcomes, just like any other event. Specifically, since Ω contains all of the possible outcomes of an experiment, the statement $P(\Omega) = 1$ tells us that *the probability that one of the possible outcomes will occur is* 1.[7]

Example 22.6 Suppose we are drawing a single card for a well-shuffled deck of 52 cards. Then the sample space contains 52 possible outcomes, and hence, from Tentative Definition 22.1, we have that

$$P(\Omega) = P(\omega_1) + \cdots + P(\omega_{52}) = 1 \tag{22.7}$$

Furthermore, since the deck is well-shuffled, the probability of each card getting drawn is equal, that is, $P(\omega_1) = \cdots = P(\omega_{52})$. Hence every card has a probability of $\frac{1}{52}$, and therefore the probability of events A, B, and C in Example 22.1 are given by

$$\begin{aligned}
P(A) &= P\left(\text{A}\clubsuit\right) = \frac{1}{52} \\
P(B) &= P\left(\left\{\text{A}\clubsuit, \text{A}\diamondsuit, \text{A}\spadesuit, \text{A}\heartsuit\right\}\right) \\
&= P\left(\text{A}\clubsuit\right) + P\left(\text{A}\diamondsuit\right) + P\left(\text{A}\spadesuit\right) + P\left(\text{A}\heartsuit\right) \\
&= 4P\left(\text{A}\clubsuit\right) = 4\left(\frac{1}{52}\right) = \frac{1}{13} \\
P(C) &= P\left(\left\{\text{A}\clubsuit, \ldots, 2\clubsuit\right\}\right) = P\left(\text{A}\clubsuit\right) + \cdots + P\left(2\clubsuit\right) \\
&= 13P\left(\text{A}\clubsuit\right) = 13\left(\frac{1}{52}\right) = \frac{1}{4}
\end{aligned}$$

[7] The creators of probability theory could have just as easily chosen the definition $P(\Omega) = 2$, in which case, we would say that *the probability that one of the possible outcomes will occur is* 2. However, we use the rule $P(\Omega) = 1$ in in order to conform with the rest of the probability literature.

Example 22.7 Suppose we are rolling a standard 6-sided die. Then the sample space contains 6 possible outcomes, and hence, from Tentative Definition 22.1, we have that

$$P(\Omega) = P\left(\text{⚀}\right) + \cdots + P\left(\text{⚅}\right) = 1 \tag{22.8}$$

Furthermore, since the probability of rolling every number is equal,

$$P\left(\text{⚀}\right) = \cdots = P\left(\text{⚅}\right) = \frac{1}{6} \tag{22.9}$$

Thus the probability of events A, B, and C in Example 22.3 are given by

$$\begin{aligned}
P(A) &= P\left(\text{⚅}\right) = \frac{1}{6} \\
P(B) &= P\left(\left\{\text{⚁}, \text{⚃}, \text{⚅}\right\}\right) = P\left(\text{⚁}\right) + P\left(\text{⚃}\right) + P\left(\text{⚅}\right) \\
&= 3\left(\frac{1}{6}\right) = \frac{1}{2} \\
P(C) &= P\left(\left\{\text{⚂}, \text{⚃}, \text{⚄}, \text{⚅}\right\}\right) \\
&= P\left(\text{⚂}\right) + P\left(\text{⚃}\right) + P\left(\text{⚄}\right) + P\left(\text{⚅}\right) \\
&= 4\left(\frac{1}{6}\right) = \frac{2}{3}
\end{aligned}$$

□

Examples 22.6 and 22.7 dealt with experiments where the sample spaces were finite and the outcomes were equally likely. In these cases, we showed that it is possible to use the properties of a probability measure to calculate the probability of an individual outcome. Furthermore, given the probabilities of individual outcomes, we were then able to calculate the probabilities of events by using (22.6). However, one could easily imagine an experiment in which the outcomes are not equally likely, as we demonstrate in the following example.

Example 22.8 Suppose that we have a die which is weighted so that ⚀ is more likely to appear than any of the other sides $\left\{ \text{⚁}, \ldots, \text{⚅} \right\}$, which are equally likely to occur, that is,

$$P\left(\text{⚀}\right) > P\left(\text{⚁}\right) = \cdots = P\left(\text{⚅}\right) \tag{22.10}$$

Then the sample space still contains the 6 possible outcomes

$$\Omega = \{\omega_1, \ldots, \omega_6\} = \left\{ \text{⚀}, \text{⚁}, \text{⚂}, \text{⚃}, \text{⚄}, \text{⚅} \right\} \tag{22.11}$$

although in this case, we do not have enough information to uniquely determine the probabilities of the individual outcomes. For instance, two feasible scenarios for the experiment are:

$$\text{Case 1)}\quad P\left(\text{⚀}\right) = 0.50, \text{ and } P\left(\text{⚁}\right) = \cdots = P\left(\text{⚅}\right) = 0.10$$

$$\text{Case 2)}\quad P\left(\text{⚀}\right) = 0.25, \text{ and } P\left(\text{⚁}\right) = \cdots = P\left(\text{⚅}\right) = 0.15$$

where the actual probabilities of the individual outcomes will depend on how heavily we weight the die.

Furthermore, note that we call Case 1) and Case 2) *feasible* scenarios since these choices for the probabilities satisfy the conditions of the probability measure as defined in Tentative Definition 22.1. In contrast, the choice $P\left(\text{⚀}\right) = 0.50$, and $P\left(\text{⚁}\right) = \cdots = P\left(\text{⚅}\right) = 0.20$, is not a valid probability measure since we would find that $P(\Omega) = 1.5$. $\square$

Example 22.8 demonstrates that if the outcomes of an experiment are not equally likely, then there are many ways of choosing the probabilities to satisfy our tentative definition of a probability measure, although the specific values will depend on the experiment being run. Furthermore, Example 22.8 shows that if you are responsible for defining the probabilities of individual outcomes, then you must be careful to ensure that they satisfy the properties of a probability measure.

Next, we consider some cases in which the sample space is infinite[8], where we use Tentative Definition 22.1 to check whether our choice of outcome probabilities meets the requirements of a probability measure.

Example 22.9 Suppose that an experiment has an infinite number of outcomes $\{\omega_1, \omega_2, \ldots\}$, where all of the outcomes $\omega_1, \omega_2, \ldots$ are equally likely, that is,

$$P(\omega_k) = a, \qquad \text{for all } k \geq 1 \tag{22.12}$$

Then the probability that one of the outcomes in Ω will occur is

$$P(\Omega) = \lim_{n\to\infty} P\left(\{\omega_1, \ldots, \omega_n\}\right) = \lim_{n\to\infty} na = \begin{cases} 0, & a = 0 \\ +\infty, & a > 0 \\ -\infty, & a < 0 \end{cases} \tag{22.13}$$

Therefore P does not satisfy property i) of Tentative Definition 22.1, and hence (22.12) is not a valid probability measure. ❒

Example 22.10 Suppose that an experiment has an infinite number of outcomes $\{\omega_1, \omega_2, \ldots\}$, where the probabilities of $\omega_1, \omega_2, \ldots$ are given by

$$P\left(\omega_k\right) \triangleq \frac{1}{k(k+1)}, \qquad \text{for } k \geq 1 \tag{22.14}$$

Then since each outcome has a positive probability, (22.14) satisfies property ii) of Tentative Definition 22.1. Furthermore, since the probability of one of the first n outcomes occurring is

$$P\left(\{\omega_1, \ldots, \omega_n\}\right) = \sum_{k=1}^{n} P\left(\omega_k\right) = 1 - \frac{1}{n+1} \tag{22.15}$$

then the probability of one of the outcomes in Ω occurring is

$$P\left(\Omega\right) = \lim_{n\to\infty} P\left(\{\omega_1, \ldots, \omega_n\}\right) = \lim_{n\to\infty} \left(1 - \frac{1}{n+1}\right) = 1 \tag{22.16}$$

Hence (22.14) also satisfies property i) of Tentative Definition 22.1. Therefore (22.14) is a valid probability measure. ❒

[8] Technically, we consider *countably infinite* sample spaces. We will come back to this issue later.

Example 22.11 Suppose that an experiment has an infinite number of outcomes $\{\omega_1, \omega_2, \ldots\}$, where the probabilities of $\omega_1, \omega_2, \ldots$ are given by

$$P(\omega_k) \triangleq \alpha^{k-1}, \qquad \text{for } k \geq 1 \tag{22.17}$$

and where α is a positive number less than 1, that is, $\alpha \in (0, 1)$. Then since each outcome has a positive probability, (22.17) satisfies property ii) of Tentative Definition 22.1. However, since the probability of one of the first n outcomes occurring is

$$P(\{\omega_1, \ldots, \omega_n\}) = \sum_{k=1}^{n} P(\omega_k) = \frac{1-\alpha^n}{1-\alpha} \tag{22.18}$$

then the probability of one of the outcomes in Ω occurring is

$$P(\Omega) = \lim_{n\to\infty} P(\{\omega_1, \ldots, \omega_n\}) = \lim_{n\to\infty} \left(\frac{1-\alpha^n}{1-\alpha}\right) = \frac{1}{1-\alpha} \neq 1 \tag{22.19}$$

Hence (22.17) does not satisfy the property i) of a probability measure. If, on the other hand, we would have defined the probabilities to be given by

$$\bar{P}(\omega_k) \triangleq (1-\alpha)\alpha^{k-1}, \qquad \text{for } k \geq 1 \tag{22.20}$$

then we would find that $\bar{P}(\Omega) = 1$. Hence (22.20) would satisfy the properties of a probability measure as given by Tentative Definition 22.1. □

Examples 22.9-22.11 demonstrate:

1) We can define probability measures over infinite sample spaces.

2) If a sample space is infinite, the outcomes cannot be equally likely.

3) Since the valid probability measure $\bar{P}$ in Example 22.11 is a function of an arbitrary real number α, there exist an infinite number of valid probability measures.

4) If we find that our first choice of probabilities results in the conclusion that $P(\Omega) = \beta$, where β is a finite positive number not equal to 1, then perhaps $\bar{P}(\omega_k) = \dfrac{P(\omega_k)}{\beta}$ is a valid probability measure.

22.4 Uncountable Sets

22.5 Sets

Since events are sets of outcomes, and Ω is the set of all possible outcomes, it follows that every event is a subset of the sample space Ω. Sets therefore play a crucial role in probability theory. Specifically, they help us answer questions such as, *What is the probability that an event does not happen?* and *What is the probability that two events both happen?* We present the idea of a subset, along with some other set relations and operations in the following definition:

Definition 22.2 Let E, F, and G be sets.

(i) If E and F contain exactly the same elements, then we say that E is *equal to* F, that is, $E = F$. Otherwise, $E \neq F$. For instance,

$$\{1, 2, 3, 4\} = \{1, 4, 2, 3\}, \qquad \{1, 3\} \neq \{1, 2, 3, 4\}$$

(ii) If all of the elements of E are also in F, and the two sets might be equal, then we say that E is a *subset* of F, that is, $E \subseteq F$. Otherwise, $E \nsubseteq F$. For instance,

$$\{1, 3\} \subseteq \{1, 3\}, \qquad \{1, 3\} \subseteq \{1, 2, 3, 4\}, \qquad \{1, 2, 3, 4\} \nsubseteq \{1, 3\}$$

(iii) If all of the elements of E are also in F, but there is at least one element in F that is not in E, then we say that E is a *proper subset* of F, that is, $E \subset F$. Otherwise, $E \not\subset F$. For instance,

$$\{1, 3\} \subset \{1, 2, 3, 4\}, \qquad \{1, 3\} \not\subset \{1, 3\}, \qquad \{1, 2, 3, 4\} \not\subset \{1, 3\}$$

(iv) If E and F are both subsets of G, and every element of G is either in E, F, or both, then we say that G is the *union* of E and F, that is, $E \cup F = G$. For instance,

$$\{1, 2, 3, 4\} \cup \{1, 3, 5, 7, 9\} = \{1, 2, 3, 4, 5, 7, 9\}$$

(v) If G is a subset of both E and F, and there are no elements common to both E and F that are not in G, then we say that G is the *intersection* of E and F, that is, $E \cap F = G$. For instance,

$$\{1, 2, 3, 4\} \cap \{1, 3, 5, 7, 9\} = \{1, 3\}$$

□

22.5.1 The Probability of an Event Not Occurring

When the sample space consists of only a small, finite number of outcomes, then one approach for calculating the probability of an event not occurring is to simply enumerate all of the other options. For instance, the probability that I do not roll a number greater than two is the probability that I roll either a one or a two. However, for more complicated problems, we need a better answer. That answer will come by way of the set complement, but before we get to that, we need to mention one of the most important sets of all, the empty set:

Definition 22.3 The empty set is the set with no elements. It is denoted by $\emptyset$. For every set E, the empty set satisfies

$$\emptyset \subseteq E, \qquad E \cup \emptyset = E, \qquad E \cap \emptyset = \emptyset \tag{22.21}$$

□

Using the empty set, we can now define the set complement:

Definition 22.4 Let E be a subset of F, that is, $E \subseteq F$. Then the *complement* of E, denoted by E^c, is the subset of F such that

$$E \cup E^c = F \quad \text{and} \quad E \cap E^c = \emptyset \tag{22.22}$$

□

A set complement requires two things: a superset F, and a subset E of that superset. In this case, the compliment E^c of E is the set of all of the elements in F that are not in E, as shown in Figure 11.

In the context of probability theory, our superset is almost always the sample space Ω. Hence the complement E^c of an event E, is the set containing all of the outcomes in Ω that are not in E. More precisely, since one of the outcomes in Ω must occur, E^c is the event that E does not occur.

Example 22.12 Let x denote the result of rolling a six-sided die. Then the sample space is

$\Omega = \Big\{$ ⚀, ⚁, ⚂, ⚃, ⚄, ⚅ $\Big\}$ (22.23)

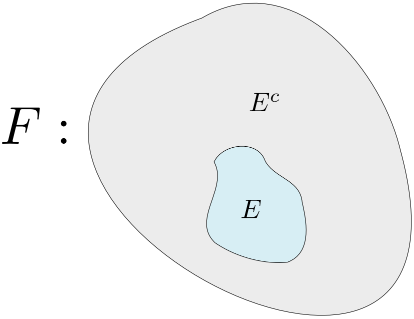


Figure 11: A subset E of F, along with the its complement E^c.

and the event that the result is greater than 2 is written as

$$E = \left\{ \text{⚂}, \text{⚃}, \text{⚄}, \text{⚅} \right\} \tag{22.24}$$

Hence the set complement of E is given by $E^c = \left\{ \text{⚀}, \text{⚁} \right\}$, that is, E^c is the event that E does not occur. □

So how does this help us? Let $\Omega = \{\omega_1, \omega_2, \ldots\}$ denote the sample space of an experiment, and let $E = \{\omega_1, \omega_4, \ldots\}$ denote an arbitrary event. Then the complement of E is given by $E^c = \{\omega_2, \omega_3, \ldots\}$, where E^c denotes the event that E does not occur. Furthermore, since

$$P(E) + P(E^c) = \Big[P(\omega_1) + P(\omega_4) + \cdots\Big] + \Big[P(\omega_2) + P(\omega_3) + \cdots\Big]$$

and $E \cup E^c = \Omega$, it follows that

$$P(E) + P(E^c) = \sum_{i=1} P(\omega_i) = P(\Omega) = 1 \tag{22.25}$$

Hence the probability that an event E does not occur is given by

$$P(E^c) = 1 - P(E) \tag{22.26}$$

22.5.2 The Probability of More Than One Event

The previous section developed a nifty little formula (22.26) relating the probability of an event E occurring to the probability of that the event E

will not occur, that is, E^c. Here we develop relations for more than one event. Specifically, let A and B denote events. Then

- $A \cap B$ denotes the set of outcomes in both A and B.
- $A \cup B$ denotes the set of outcomes in either A, B, or both A and B.

Therefore

- The probability that *both* A *and* B occur is the probability that one of the outcomes in $A \cap B$ occurs, that is, $P(A \cap B)$.
- The probability that *either* A occurs, B occurs, *or* both A and B occur is the probability that one of the outcomes in $A \cup B$ occurs, that is, $P(A \cup B)$.

Example 22.13 Let x denote the result of rolling a six-sided die. Furthermore, as in Example 22.3, let A denote the event that the roll will be a six, and let B denote the event that the roll will be even. Then the probability that both A and B occur is given by

$$P\left(A \cap B\right) = P\left(\left\{\text{⚅}\right\} \bigcap \left\{\text{⚁}, \text{⚃}, \text{⚅}\right\}\right) = P\left(\text{⚅}\right)$$

Furthermore, the probability that either A, B, or both occur is

$$\begin{aligned} P\left(A \cup B\right) &= P\left(\left\{\text{⚅}\right\} \bigcup \left\{\text{⚁}, \text{⚃}, \text{⚅}\right\}\right) \\ &= P\left(\left\{\text{⚁}, \text{⚃}, \text{⚅}\right\}\right) \\ &= P\left(\text{⚁}\right) + P\left(\text{⚃}\right) + P\left(\text{⚅}\right) \end{aligned}$$

Next, suppose that A and B have no common outcomes. Then

$$P\left(A \cap B\right) = P\left(\emptyset\right) \tag{22.27}$$

where, since $\emptyset = \Omega^c$, from (22.26) we have that

$$P\left(\emptyset\right) = 1 - P\left(\Omega\right) = 1 - 1 = 0 \tag{22.28}$$

Hence if A and B have no common outcomes, then

$$P(A \cap B) = P(\emptyset) = 0 \tag{22.29}$$

Furthermore, since A and B have no common outcomes, we can write A and B in the form $A = \{\omega_1, \ldots, \omega_n\}$ and $B = \{\omega_{n+1}, \ldots, \omega_p\}$. Therefore the probability of either A or B occurring is

$$\begin{aligned} P(A \cup B) &= P(\{\omega_1, \ldots, \omega_n, \omega_{n+1}, \ldots, \omega_p\}) && (22.30) \\ &= P(\omega_1) + \cdots + P(\omega_n) + P(\omega_{n+1}) + \cdots + P(\omega_p) && (22.31) \\ &= P(A) + P(B) && (22.32) \end{aligned}$$

Remark 22.1 Trying to compute the probability of an empty set may sound a bit strange. The easiest way to understand it is to think of $P(\emptyset)$ as *the probability that none of the outcomes in* Ω *occurs.* Specifically, since one of the outcomes in Ω must occur as the result of the experiment, the event $\emptyset$ is impossible, which is reflected in the fact that $P(\emptyset) = 0$. □

Finally, suppose that A and B contain at least one common outcome, that is,

$$A \cap B \neq \emptyset \tag{22.33}$$

then we can write A and B in the form $A = \{\omega_1, \ldots, \omega_n, \omega_{n+1}, \ldots, \omega_\ell\}$ and $B = \{\omega_{n+1}, \ldots, \omega_\ell, \omega_{\ell+1}, \ldots, \omega_p\}$, where $\{\omega_{n+1}, \ldots, \omega_\ell\}$ is the set of outcomes common to both A and B. Hence

$$\begin{aligned} P(A) + P(B) =& P(\omega_1) + \cdots + P(\omega_n) \\ & + 2P(\omega_{n+1}) + \cdots + 2P(\omega_\ell) \\ & + P(\omega_{\ell+1}) + \cdots + P(\omega_p) \\ =& P(\{\omega_1, \ldots, \omega_p\}) + P(\{\omega_{n+1}, \ldots, \omega_\ell\}) \\ =& P(A \cup B) + P(A \cap B) \end{aligned}$$

which is usually written in the form

$$P(A \cup B) = P(A) + P(B) - P(B \cap A) \tag{22.34}$$

Example 22.14 Let x denote the result of rolling a six-sided die. Furthermore, as in Example 22.3, let A denote the event that the roll will be a six, and let B denote the event that the roll will be even. Then the probability that either A, B, or both occur is

$$P(A \cup B) = P\left(\left\{\text{⚅}\right\} \cup \left\{\text{⚁}, \text{⚃}, \text{⚅}\right\}\right)$$

where, from (22.34), we find that

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= P\left(\text{⚅}\right) + P\left(\left\{\text{⚁}, \text{⚃}, \text{⚅}\right\}\right) - P\left(\text{⚅}\right) \\ &= P\left(\left\{\text{⚁}, \text{⚃}, \text{⚅}\right\}\right) \end{aligned}$$

which agrees with what we found in Example 22.12. □

Taking this one step further, we can ask:

Question 1) If event A occurs, what is the probability that event B also occurs?

To answer this question, let's first examine the seemingly stupid question:

Question 2) If event A occurs, what is the probability that event A also occurs?

If I do not know whether event A occurs, then the answer to Question 2) is simple: $P(A)$ since the question could be shortened to *What is the probability that event A also occurs?*

first note that this is different asking whether an outcome in both A and B $(A \cap B)$ occurs since, in this case, we are restricting our attention to

If we know that A has occurred, then this is like asking, *What is the probability that one of the events in B which overlaps A occurs*, that is,

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \tag{22.35}$$

To answer this question, let's go

Definition 22.5 *conditional probability of A given B*

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{22.36}$$

□

Definition 22.6 Let $(\Omega, \mathcal{S}, P)$ denote a probability space, where A and B are two events in the σ-algebra $\mathcal{S}$, that is, $A, B \in \mathcal{S}$. Then A and B are *independent* if

$$P\big(A \cap B\big) = P\big(A\big)P\big(B\big) \tag{22.37}$$

□

22.6 Countably Infinite

When would I have an infinite sample space? For instance, if I ask you to choose a positive integer, then there exist an infinite number of possible outcomes since the set of integers is infinite.

TODO: Show that for uncountably infinite sets, there always exists an event which kills our probability measure. However, if we restrict our attention to just particular subsets of the sample space, as defined by the σ-algebra, then we will never have any contradictions.

22.7 Probability Spaces

Definition 22.7 Let Ω denote the possibly infinite set $\{\omega_1, \omega_2, \ldots\}$. Furthermore, let $\mathcal{S}$ denote a set of subsets of Ω, that is, for every $A \in \mathcal{S}$, $A \subseteq \Omega$. Then we have the following definitions:

(i) If, for all $A, B \in \mathcal{S}$,

$$A^c \in \mathcal{S} \quad \text{and} \quad A \cup B \in \mathcal{S} \tag{22.38}$$

then $\mathcal{S}$ is called an *algebra on* Ω, or simply *an algebra.*

(ii) If $\mathcal{S}$ is an algebra, and for all $A_1, A_2, \ldots \in \mathcal{S}$,

$$\bigcup_{i=1}^{\infty} A_i \in \mathcal{S} \tag{22.39}$$

then $\mathcal{S}$ is called a *σ-algebra on* Ω, or simply a *σ-algebra.*

□

Fact 22.1 If $\mathcal{S}$ denotes an algebra on Ω, then:

(i) $\Omega \in \mathcal{S}$

(ii) $\emptyset \in \mathcal{S}$

(iii) For all $A, B \in \mathcal{S}$, $A \cap B \in \mathcal{S}$.

Proof

(i) Let $B \triangleq A^c$. Since $\mathcal{S}$ is an algebra, then from (22.38) it follows that $B = A^c \in \mathcal{S}$ and $A \cup B = A \cup A^c = \Omega \in \mathcal{S}$.

(ii) From (i), we have that $\Omega \in \mathcal{S}$. Furthermore, since $\mathcal{S}$ is an algebra, then from (22.38) it follows that $\Omega^c = \emptyset \in \mathcal{S}$

(iii) Since $\mathcal{S}$ is an algebra, then from (22.38) it follows that $A^c, B^c \in \mathcal{S}$ for all $A, B \in \mathcal{S}$. Hence $A^c \cup B^c \in \mathcal{S}$ and $\big(A^c \cup B^c\big)^c \in \mathcal{S}$. Finally, since $A \cap B = \big(A^c \cup B^c\big)^c$, it follows that $A \cap B \in \mathcal{S}$

□

Definition 22.8 Let Ω be a set. Then the *power set* $P(\Omega)$ is the set of all subsets of Ω, including $\emptyset$ and Ω. □

Fact 22.2 If the sample space Ω has a finite number of elements, then the power set of Ω, $P(\Omega)$ is a σ-algebra.

Proof TODO □

Remark 22.2 Apparently Cantor's theorem shows that if the sample space Ω has a countably infinite number of elements, then the power set is uncountably infinite. □

Example 22.15 Let the sample space be given by $\Omega \triangleq \{1, 2, 3\}$. Then the power set of Ω is given by

$$P(\Omega) = \big\{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\big\} \tag{22.40}$$

where $P(\Omega)$ is a σ-algebra. □

Definition 22.9 Let Ω denote the sample space of an experiment. Also, let $\mathcal{S}$ be a σ-algebra on Ω. Then P is a *probability measure* for the σ-algebra $\mathcal{S}$ if all of the following hold:

(i) $P(\Omega) = 1$

(ii) For every event $A \in \mathcal{S}$, $P(A) \geq 0$.

(iii) For all mutually disjoint $A_1, A_2, \ldots \in \mathcal{S}$,

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P\left(A_i\right) \tag{22.41}$$

where, by mutually disjoint, we mean that for all $i \geq 1$ and $j \geq 1$ such that $i \neq j$, $A_i \cap A_j = \emptyset$.

Furthermore, the triple $(\Omega, \mathcal{S}, P)$ is called a *probability space.* □

SECTION 23

Random Variables

Wait a second, you're telling me that we spent all of this time and energy developing probability theory, and we can't even define a probability space over the real numbers? Well, yes and no... Of course, you could always use the Borel algebra to define a probability space over all of the subsets of the reals bounded by rational numbers, but that's not what you want, is it? Fortunately, there is a way forward, and that is to define a function called a random variable.

Definition 23.1 Let $x, y \in \mathbb{R}^n$, where

$$x \triangleq \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \qquad y \triangleq \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \tag{23.1}$$

If $x_i \leq y_i$ for all $i \in [1, n]$, then we write $x \leq y$.

Definition 23.2 Let $(\Omega, \mathcal{S}, P)$ denote a probability space, and let X be a function which maps the sample space Ω into $\mathbb{R}^n$, that is, $X : \Omega \rightarrow \mathbb{R}^n$. Furthermore, let $x \in \mathbb{R}^n$, and let $S(x) \subseteq \Omega$ denote the set of all outcomes $\omega \in \Omega$ for which $X(\omega) \leq x$, that is,

$$S(x) = \{\omega : X(\omega) \leq x\} \subseteq \Omega \tag{23.2}$$

- If $S(x) \in \mathcal{S}$ for all $x \in \mathbb{R}^n$, then X is called a *random variable.*
- If X is a *random variable*, then

  $$F(x) = P\big(S(x)\big) \tag{23.3}$$

  is called the *cumulative distribution function* of X, which for convenience, we write as

  $$F(x) = P\left(X \leq x\right) \tag{23.4}$$

Remark 23.1 When X is a vector-random variable ($X : \Omega \to \mathbb{R}^n$ where $n > 1$), the cumulative distribution function $F(x)$ still evaluates to a scalar, despite the fact that $x \in \mathbb{R}^n$. In this case, letting $X_i(\omega)$ and x_i denote the i^{th} entries of $X(\omega)$ and x, respectively, the cumulative distribution function $F(x)$ is interpreted as the set of outcomes in the σ-algebra $\mathcal{S}$ such that $X_i(\omega) \leq x_i$ for all $i = 1, \dots, n$, that is,

$$\begin{aligned} F(x) &= P\Big(S(x)\Big) = P\Big(\{\omega : X(\omega) \leq x\}\Big) \\ &= P\Big(\big\{\omega : X_1(\omega) \leq x_1, \text{ and } X_2(\omega) \leq x_2, \cdots, \text{ and } X_n(\omega) \leq x_n\big\}\Big) \end{aligned}$$

□

Remark 23.2 A lot of confusion stems from the fact that people think of a random variable as either a real number, an outcome, or an event. It is none of these. It is a function which maps the outcomes of an experiment to real numbers, that is, $X : \Omega \to \mathbb{R}^n$. □

Remark 23.3 The probability measure P is only defined to operate on elements of the σ-algebra $\mathcal{S}$, that is, sets of outcomes of the experiment. For instance, if $E \in \mathcal{S}$ is an event composed of the outcomes $\{\omega_1, \dots, \omega_n\}$, then the probability that the event E occurs is the probability that one of the outcomes in E occurs, that is,

$$P(E) = P\Big(x = \omega_1, \quad \text{or} \quad x = \omega_2, \quad \dots, \quad \text{or} \quad x = \omega_n\Big) \tag{23.5}$$

One example of this is given by (23.3), where $S(x)$ is an event composed of all of the outcomes ω for which the random variable $X(\omega) \in \mathbb{R}^n$ is less than or equal to $x \in \mathbb{R}^n$. On the other hand, (23.4) is not valid in any mathematical sense, it is strictly a shorthand notation for (23.3). □

Fact 23.1 For every probability space $(\Omega, \mathcal{S}, P)$, the constant function

$$X(\omega) = \alpha, \qquad \text{for all } \omega \in \Omega \tag{23.6}$$

is a random variable.

Proof Since $\mathcal{S}$ is a σ-algebra, then from Fact 22.1 it follows that $\emptyset, \Omega \in \mathcal{S}$. Hence

$$\begin{aligned} &\text{for all } x < \alpha : \quad \{\omega : X(\omega) \leq x\} = \emptyset \subseteq \mathcal{S} \\ &\text{for all } x \geq \alpha : \quad \{\omega : X(\omega) \leq x\} = \Omega \subseteq \mathcal{S} \end{aligned} \tag{23.7}$$

□

Example 23.1 Let $(\Omega, \mathcal{S}, P)$ denote a probability space where

$$\Omega = \{1,2,3,4\} \tag{23.8}$$

$$\mathcal{S} = \Big\{\emptyset, \{1,2\}, \{3,4\}, \Omega\Big\} \tag{23.9}$$

Then the function

$$X(\omega) = \omega \tag{23.10}$$

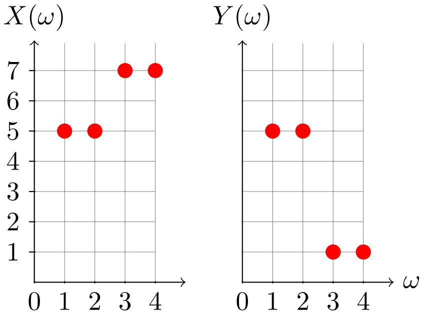
is not a random variable since $A(1.5)$ (the set of outcomes for which $X(\omega) \leq 1.5$) is not in the σ-algebra $\mathcal{S}$, that is,

$$A(1.5) = \{\omega : X(\omega) \leq 1.5\} = \{1\} \notin \mathcal{S} \tag{23.11}$$

However, the functions

$$X(\omega) = \begin{cases} 5, & \omega \in \{1,2\} \\ 7, & \omega \in \{3,4\} \end{cases}$$

$$Y(\omega) = \begin{cases} 5, & \omega \in \{1,2\} \\ 1, & \omega \in \{3,4\} \end{cases}$$



are both random variables since

$$\begin{array}{rl}
\text{for all } x < 1: & \left\{\begin{array}{l} \{\omega : X(\omega) \leq x\} = \emptyset \in \mathcal{S} \\ \{\omega : Y(\omega) \leq x\} = \emptyset \in \mathcal{S} \end{array}\right. \\ \hline
\text{for all } 1 \leq x < 5: & \left\{\begin{array}{l} \{\omega : X(\omega) \leq x\} = \emptyset \in \mathcal{S} \\ \{\omega : Y(\omega) \leq x\} = \{3,4\} \in \mathcal{S} \end{array}\right. \\ \hline
\text{for all } 5 \leq x < 7: & \left\{\begin{array}{l} \{\omega : X(\omega) \leq x\} = \{1,2\} \in \mathcal{S} \\ \{\omega : Y(\omega) \leq x\} = \{1,2,3,4\} = \Omega \in \mathcal{S} \end{array}\right. \\ \hline
\text{for all } x \geq 7: & \left\{\begin{array}{l} \{\omega : X(\omega) \leq x\} = \{1,2,3,4\} = \Omega \in \mathcal{S} \\ \{\omega : Y(\omega) \leq x\} = \{1,2,3,4\} = \Omega \in \mathcal{S} \end{array}\right.
\end{array}$$

□

Example 23.2 Let $(\Omega, \mathcal{S}, P)$ denote a probability space where

$$\Omega = \big(0,4\big] \tag{23.12}$$

$$\mathcal{S} = \Big\{\emptyset, \big(0,2\big], \big(2,4\big], \Omega\Big\} \tag{23.13}$$

Then the function

$$X(\omega) = \omega \tag{23.14}$$

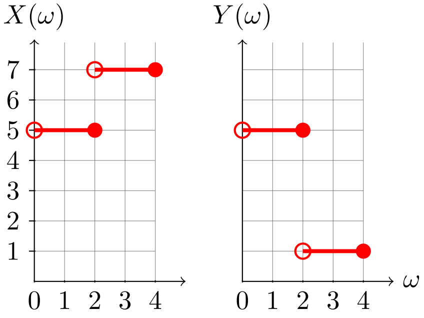
is not a random variable since $A(1)$ (the set of outcomes for which $X(\omega) \leq 1$) is not in the σ-algebra $\mathcal{S}$, that is,

$$A(1) = \{\omega : X(\omega) \leq 1\} = \big(0,1\big] \notin \mathcal{S} \tag{23.15}$$

However, the functions

$$X(\omega) = \begin{cases} 5, & 0 < \omega \leq 2 \\ 7, & 2 < \omega \leq 4 \end{cases}$$

$$Y(\omega) = \begin{cases} 5, & 0 < \omega \leq 2 \\ 1, & 2 < \omega \leq 4 \end{cases}$$



are both random variables since

$$\text{for all } x < 1 : \quad \begin{cases} \{\omega : X(\omega) \leq x\} = \emptyset \in \mathcal{S} \\ \{\omega : Y(\omega) \leq x\} = \emptyset \in \mathcal{S} \end{cases}$$

$$\text{for all } 1 \leq x < 5 : \quad \begin{cases} \{\omega : X(\omega) \leq x\} = \emptyset \in \mathcal{S} \\ \{\omega : Y(\omega) \leq x\} = \big(2,4\big] \in \mathcal{S} \end{cases}$$

$$\text{for all } 5 \leq x < 7 : \quad \begin{cases} \{\omega : X(\omega) \leq x\} = \big(0,2\big] \in \mathcal{S} \\ \{\omega : Y(\omega) \leq x\} = \big(0,4\big] = \Omega \in \mathcal{S} \end{cases}$$

$$\text{for all } x \geq 7 : \quad \begin{cases} \{\omega : X(\omega) \leq x\} = \big(0,4\big] = \Omega \in \mathcal{S} \\ \{\omega : Y(\omega) \leq x\} = \big(0,4\big] = \Omega \in \mathcal{S} \end{cases}$$

23.1 The Probability Density Function

Definition 23.3 Let $(\Omega, \mathcal{S}, P)$ denote a probability space, and let $F : \mathbb{R}^n \to [0, 1]$ denote the cumulative distribution function of a random variable $X : \Omega \to \mathbb{R}^n$ on that probability space. If there exists a positive integrable function $f : \mathbb{R}^n \to \mathbb{R}$ such that for all $x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n$,

$$F(x) = P(X \leq x) = \int\limits_{-\infty}^{x_1} \cdots \int\limits_{-\infty}^{x_n} f(\tau) d\tau_1 \ldots d\tau_n \tag{23.16}$$

then f is called the *probability density function* of X. Specifically, if F is differentiable, then

$$f(x) \triangleq \frac{\partial^n F(x)}{\partial x_1 \cdots \partial x_n} \tag{23.17}$$

□

Remark 23.4 If X is a scalar random variable, that is, $X : \Omega \to \mathbb{R}$, then the probability density function f of X satisfies

$$F(x) = P(X \leq x) = \int\limits_{-\infty}^{x} f(\tau) d\tau \tag{23.18}$$

□

Although a random variable X maps $\Omega \to \mathbb{R}^n$, we will typically drop the function notation and say things like: *Consider the random variable* $X \in \mathbb{R}^n$, in which case, we just want to emphasize that the range of the random variable is n-dimensional. This is, of course, a blatant abuse of notation. However, we will rarely need to refer to the domain Ω directly.

Instead, we will typically define a random variable by its probability density function. For instance, we will say things like: *Consider the Gaussian random variable* $X \in \mathbb{R}^n$, in which case, the word *Gaussian* tells us the structure of its probability density function.

Next, we introduce some of the common random variables encountered in the literature, where instead of introducing the mapping $X : \Omega \to \mathbb{R}^n$ directly, we define random variables in terms of their probability density functions.

Definition 23.4 Let $\mu, \sigma \in \mathbb{R}$, where $\sigma > 0$. If $X \in \mathbb{R}$ is a scalar random variable with the probability density function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \tag{23.19}$$

then X is called a *Gaussian random variable*, and $f(x)$ is called the *normal distribution.*

Definition 23.5 Let $\alpha, \beta \in \mathbb{R}$, where $\alpha > 0$ and $\beta > 0$. Also, let $\Gamma(\alpha)$ denote the gamma function, that is,

$$\Gamma(\alpha) \triangleq \int\limits_0^\infty x^{\alpha-1} \exp\left(-x\right) dx \tag{23.20}$$

If $X \in \mathbb{R}$ is a scalar random variable with the probability density function

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} \exp\left(-\frac{x}{\beta}\right) \tag{23.21}$$

then X is called a *gamma-distributed random variable*, and $f(x)$ is called the *gamma distribution.*

TODO: Uncomment "Sums and Functions of Random Variables"

23.2 Functions of a Random Variable and the Expected Value

Give the interpretation of what it means to have a function of a random variable. After this, we could define the distribution and density functions of a function of a random variable. Finally, we would explain the expected value, and describe intuitively what it is accomplishing. What is missing is a relationship between the expected value and the Probability function.

Definition 23.6 Let $f : \mathbb{R}^n \to \mathbb{R}$ denote the probability density function of the random variable $X : \Omega \to \mathbb{R}^n$, and let $g : \mathbb{R}^n \to \mathbb{R}^{p \times m}$. Then

$$\mathbb{E}\big[g(X)\big] \triangleq \int\limits_{-\infty}^{\infty} \cdots \int\limits_{-\infty}^{\infty} g(\tau) f(\tau) d\tau_1 \dots d\tau_n \tag{23.22}$$

is called the *expected value of* $g(X)$. Furthermore:

(i) $\mathbb{E}\big[X\big]$ is called the *mean of* X.

(ii) $\mathbb{E}\Big[\Big(X - \mathbb{E}[X]\Big)\Big(X - \mathbb{E}[X]\Big)^T\Big]$ is called the *covariance of* X.

(iii) If X is scalar, then $\mathbb{E}\big[(X - \mathbb{E}[X])^2\big]$ is called the *variance of* X.

□

Remark 23.5 If X is a scalar random variable, that is, $X : \Omega \to \mathbb{R}$, then the expected value of $g(X)$ is given by

$$\mathbb{E}\big[g(X)\big] \triangleq \int\limits_{-\infty}^{\infty} g(\tau) f(\tau) d\tau \tag{23.23}$$

□

Example 23.3 Let $X \in \mathbb{R}$ be a Gaussian random variable with the probability density function (23.19). Then the mean of X is given by

$$\mathbb{E}[X] = \int\limits_{-\infty}^{\infty} x f(x) dx = \frac{1}{\sqrt{2\pi\sigma^2}} \int\limits_{-\infty}^{\infty} x \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] dx \tag{23.24}$$

Hence letting $y \triangleq (x-\mu)/\sigma$, we find that

$$\mathbb{E}[X] = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \left(\sigma y + \mu \exp\left[-\frac{1}{2}y^2\right]\right) \sigma dy \tag{23.25}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \mu \exp\left[-\frac{1}{2}y^2\right] \sigma dy = \frac{\mu}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}y^2\right] dy \tag{23.26}$$

where, using an integral table, we find that

$$\mathbb{E}[X] = \mu \tag{23.27}$$

Furthermore, the variance of X is given by

$$\mathbb{E}\left[(X-\mu)^2\right] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx \tag{23.28}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} (x-\mu)^2 \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] dx \tag{23.29}$$

$$= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y^2 \exp\left[-\frac{1}{2}y^2\right] dy = \sigma^2 \tag{23.30}$$

❒

Example 23.4 Let $X \in \mathbb{R}$ be a gamma-distributed random variable with the probability density function (23.21). Then the mean of X is given by

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xf(x)dx = \int_{-\infty}^{\infty} \frac{x^\alpha}{\beta^\alpha \Gamma(\alpha)} \exp\left(-\frac{x}{\beta}\right) dx \tag{23.31}$$

Hence letting $y \triangleq x/\beta$, we find that

$$\mathbb{E}[X] = \frac{\beta}{\Gamma(\alpha)} \int_{-\infty}^{\infty} y^\alpha \exp(-y)\, dy = \frac{\beta\Gamma(\alpha+1)}{\Gamma(\alpha)} = \frac{\beta\alpha\Gamma(\alpha)}{\Gamma(\alpha)} = \beta\alpha \tag{23.32}$$

Furthermore, the variance of X is given by

$$\mathbb{E}\left[\left(X-\beta\alpha\right)^2\right] = \int\limits_{-\infty}^{\infty} \left(x-\beta\alpha\right)^2 \frac{x^{\alpha-1}}{\beta^{\alpha}\Gamma(\alpha)} \exp\left(-\frac{x}{\beta}\right) dx = \beta^2\alpha \tag{23.33}$$

❒

23.3 Multiple Random Variables

Given a probability space $(\Omega, \mathcal{S}, P)$, we can of course construct more than one random variable, in which case, we need to introduce the concept of the joint distribution of two random variables:

Definition 23.7 Let $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^p$ be random variables defined on the probability space $(\Omega, \mathcal{S}, P)$. Then the probability that $X \leq x$ *and* $Y \leq y$ is called the *joint distribution of* X *and* Y, which we denote by

$$F(x,y) \triangleq P\Big(\big(X \leq x\big) \cap \big(Y \leq y\big)\Big) \tag{23.34}$$

Furthermore, if there exists a positive integrable function $f : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$ such that for all $x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n$ and $y = \begin{bmatrix} y_1 \\ \vdots \\ y_p \end{bmatrix} \in \mathbb{R}^p$,

$$F(x,y) = \int\limits_{-\infty}^{x_1} \cdots \int\limits_{-\infty}^{x_n} \int\limits_{-\infty}^{y_1} \cdots \int\limits_{-\infty}^{y_p} f(\tau, s) ds_1 \ldots ds_p d\tau_1 \ldots d\tau_n \tag{23.35}$$

then $f(x,y)$ is called the *joint probability density function of* X *and* Y. Specifically, if $F(x,y)$ is differentiable, then

$$f(x,y) = \frac{\partial^{n+p} F(x,y)}{\partial x_1 \cdots \partial x_n \partial y_1 \cdots \partial y_p} \tag{23.36}$$

□

Remark 23.6 If the random variables X and Y are scalar, then

$$f(x,y) = \frac{\partial^2 F(x,y)}{\partial x \partial y} \tag{23.37}$$

□

Remark 23.7 The joint distribution of two random variables $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^p$ behaves exactly like the cumulative distribution function of a random vector $Z = \begin{bmatrix} X \\ Y \end{bmatrix}$. The different name is simply a matter of perspective. Most classical textbooks start by introducing random variables as scalars, in which case the joint distribution extends random variables to more than one dimension. □

Remark 23.8 Although it is not always apparent, when you consider statements of two random variables, they are almost always defined on the same probability space.

Definition 23.8 Let $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^p$ be random variables defined on the probability space $(\Omega, \mathcal{S}, P)$, where

(i) $f(x, y)$ denotes the joint probability density function of X and Y.

(ii) $F(x, y)$ denotes the joint distribution of X and Y.

Then the *marginal probability density function of* X is given by

$$f_X(x) \triangleq \int\limits_{-\infty}^{\infty} \cdots \int\limits_{-\infty}^{\infty} f(x, y) dy_1 \dots dy_p \tag{23.38}$$

and the *marginal distribution of* X is given by

$$F_X(x) \triangleq F(x, \infty) = P\big(X \leq x\big) = \int\limits_{-\infty}^{x_1} \cdots \int\limits_{-\infty}^{x_n} f_X(\tau) d\tau_1 \dots d\tau_n \tag{23.39}$$

$$= \int\limits_{-\infty}^{x_1} \cdots \int\limits_{-\infty}^{x_n} \left[\int\limits_{-\infty}^{\infty} \cdots \int\limits_{-\infty}^{\infty} f(\tau, y) dy_1 \dots dy_p \right] d\tau_1 \dots d\tau_n \tag{23.40}$$

Remark 23.9 If you compare the cumulative distribution function (23.16) with the marginal distribution function (23.39), you will see that they are very similar form. In fact, they are equal; they both represent the probability that $X \leq x$. So why define the marginal distribution? The reason is twofold:

(1) The word *marginal* hints to the reader that there are other random variables living on our probability space $(\Omega, \mathcal{S}, P)$.

(2) Equation (23.39) explicitly tells us how to derive the marginal distribution functions from the joint distribution function. On the other hand, we cannot derive the joint distribution function of two random variables X and Y from the probability density functions $f_X(x)$ and $f_Y(y)$ of the random variables. The reason we cannot do so is because $f_X(x)$ and $f_Y(y)$ yield no information about how the random variables X and Y interact, which is precisely what the joint distribution describes.

□

Example 23.5 Suppose we want to run two different experiments, where Ω_1 and Ω_2 denote the sample spaces of the first and second experiments, respectively, and where X is a random variable for the first experiment, and Y is a random variable for the second experiment. Then of course, we can make probabilistic statements about each of the experiments independently. This is called the marginal distribution.

However, if we want to make a statement about the result of both experiments together, then we need to pretend that we just ran one experiment, in which case, we talk about the joint distribution. □

Next, we will introduce the concept of independence for random variables:

Definition 23.9 Let $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^p$ be random variables defined on the probability space $(\Omega, \mathcal{S}, P)$. Furthermore, let

- $S(x) \subseteq \Omega$ denote the set of all outcomes $\omega \in \Omega$ for which $X(\omega) \leq x$
- $B(x) \subseteq \Omega$ denote the set of all outcomes $\omega \in \Omega$ for which $Y(\omega) \leq y$

Then we say that X *is independent of* Y if, for all $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^p$, $S(x)$ and $B(y)$ are independent events, that is,

$$P\Big(S(x) \cap B(y)\Big) = P\Big(S(x)\Big)P\Big(B(y)\Big) \tag{23.41}$$

which for convenience, we write as

$$P\Big(\big(X \leq x\big) \cap \big(Y \leq y\big)\Big) = P\Big(X \leq x\Big)P\Big(Y \leq y\Big) \tag{23.42}$$

□

Fact 23.2 Let $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^p$ be independent random variables, where

- $F(x, y)$ denotes the joint distribution of X and Y
- $F_X(x)$ denotes the marginal distribution of X
- $F_Y(y)$ denotes the marginal distribution of Y
- $f(x, y)$ denotes the joint probability density function of X and Y

- $f_X(x)$ denotes the marginal probability density function of X
- $f_Y(y)$ denotes the marginal probability density function of Y

Then for all $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^p$,

$$F(x,y) = F_X(x)F_Y(y) \tag{23.43}$$

$$f(x,y) = f_X(x)f_Y(y) \tag{23.44}$$

Proof TODO □

Definition 23.10 Let X and Y be random variables defined on the probability space $(\Omega, \mathcal{S}, P)$, where $f(x,y)$ denotes the joint probability density function of X and Y. Then the *expected value* of a function $g(X,Y)$ is denoted by

$$\mathbb{E}\Big[g(X,Y)\Big] \triangleq \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)dxdy \tag{23.45}$$

□

Example 23.6 Let X and Y denote two random variables. Then the expected value of $X+Y$ is given by

$$\begin{aligned}
\mathbb{E}\Big[X+Y\Big] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \Big(x+y\Big)f(x,y)dxdy \\
&= \int_{-\infty}^{\infty} x\left(\int_{-\infty}^{\infty} f(x,y)dy\right)dx + \int_{-\infty}^{\infty} y\left(\int_{-\infty}^{\infty} f(x,y)dx\right)dy \\
&= \int_{-\infty}^{\infty} xf_X(x)dx + \int_{-\infty}^{\infty} yf_Y(y)dy \\
&= E\big[X\big] + E\big[Y\big]
\end{aligned}$$

□

Definition 23.11 Let X and Y denote random variables on the sample space $(\Omega, \mathcal{S}, P)$, where $X : \Omega \to \mathbb{R}^n$ and $Y : \Omega \to \mathbb{R}^p$. Then the *covariance* of X and Y, denoted by $\operatorname{cov}(X,Y)$, is the expected value

$$\operatorname{cov}(X,Y) \triangleq \mathbb{E}\left[\Big(X - \mathbb{E}\big[X\big]\Big)\Big(Y - \mathbb{E}\big[Y\big]\Big)^H\right] \tag{23.46}$$

Furthermore, if $\mathrm{cov}(X, Y) = 0$, then we say that X and Y are *uncorrelated*. Otherwise, they are *correlated*. □

Remark 23.10 If X and Y are multivariate random variables, that is, either $n > 1$ or $p > 1$, then (23.46) is sometimes called the *covariance matrix* to emphasize the fact that the random variables are not scalar. □

23.4 Multivariable Random Variables

Definition 23.12 Let $\mu \in \mathbb{R}^n$ and let $P \in \mathbb{R}^{n \times n}$ be a symmetric positive definite matrix. If $X \in \mathbb{R}^n$ is a random variable with the probability density function

$$f(x) = \frac{1}{\sqrt{(2\pi)^n \mathtt{det}[P]}} \exp\left[-\frac{1}{2}\big(x - \mu\big)^T P^{-1} \big(x - \mu\big)\right] \tag{23.47}$$

then X is called a *multivariate Gaussian random variable*, and $f(x)$ is called the *multivariate normal distribution*, although sometimes we drop the word *multivariate* for conciseness. □

Example 23.7 Let X be a multivariate Gaussian random variable with the probability density function (23.47). Then the mean of X is given by

$$\mathbb{E}[X] = \int\limits_{-\infty}^{\infty} x f(x) dx \tag{23.48}$$

$$= \frac{1}{\sqrt{(2\pi)^n \mathtt{det}[P]}} \int\limits_{-\infty}^{\infty} x \exp\left[-\frac{1}{2}\big(x - \mu\big)^T P^{-1} \big(x - \mu\big)\right] dx \tag{23.49}$$

□

Example 23.8 Let $X_1, \ldots, X_n$ denote a sequence of independent and identically distributed Gaussian random variables with mean $\mu = 0$ and variance $\sigma^2 = 1$. Then the expected value of the sum $X_1^2 + \cdots + X_n^2$ is given by

$$\mathbb{E}\Big[X_1^2 + \cdots + X_n^2\Big] = \mathbb{E}\Big[X_1^2\Big] + \cdots + \mathbb{E}\Big[X_n^2\Big] = n \tag{23.50}$$

Furthermore, for all $r \geq 1$,

$$\mathbb{E}\left[\left(X_1^2 + \cdots + X_n^2 - n\right)^r\right] = \tag{23.51}$$

Hence, we see that the sequence behaves like a gamma-distributed random variable, where $\beta =$ and α. In fact, this is exactly the case. One can show that

SECTION 24

Random Processes

24.1 A Random Walk

Suppose we are out for an aimless stroll on a Saturday afternoon. We have no plans, and no place to be, so our walk is truly random. Hence our position $y(k)$ at some time k o'clock is related to our position at time $(k+1)$ o'clock by

$$y(k+1) = y(k) + w(k) \tag{24.1}$$

where $w(k)$ denotes a random number. For instance, we might move $w(k) = 1\text{km}$ between times k and $k+1$, or we might not move at all, in which case, $y(k+1) = y(k)$ and $w(k) = 0$. Furthermore, if we repeat this process all day, then given our initial position $y(k_0)$ at the start of the day, our position a few hours later is given by

$$y(k_0+n) = y(k_0) + \sum_{i=0}^{n-1} w(k_0+i) \tag{24.2}$$

where $w(k_0), \ldots, w(k_0+n-1)$ denote all of the random movements we've made throughout the day. The important point is that both of the sequences $w(k_0), \ldots, w(k_0+n-1)$ and $y(k_0), \ldots, y(k_0+n)$ are *random processes*:

Definition 24.1 Let T denote a set of times. If $X(t)$ is a random variable for all $t \in T$, then X is called a *random process*. More formally, let $(\Omega, \mathcal{S}, P)$ denote a probability space, and let X be a function which maps time and the sample space Ω into $\mathbb{R}^n$, that is, $X : T \times \Omega \to \mathbb{R}^n$. Furthermore, let $A(t,x) \subseteq \Omega$ denote the set of all outcomes $\omega \in \Omega$ for which $X(t,\omega) \leq x$, that is,

$$A(t,x) = \{\omega : X(t,\omega) \leq x\} \subseteq \Omega \tag{24.3}$$

Then X is called a *random process* if $A(t,x) \in \mathcal{S}$ for all $t \in T$ and $x \in \mathbb{R}^n$. □

Remark 24.1 In the continuous-time case, the set of times T is usually a subset of $\mathbb{R}$. in the discrete-time case, T is usually a subset of $\mathbb{Z}$. □

Remark 24.2 The notation $X(t)$ is just a convenient shorthand for $X(t, \omega)$ since we rarely need to reference the sample space Ω directly. This notation is in keeping with our notation for random variables. □

Remark 24.3 You can think of a random process as a signal X, where at each time t, $X(t)$ is a random variable. □

Remark 24.4 Random processes are sometimes called *ensembles.* Discrete random processes are sometimes called *random sequences.* □

Example 24.1 Let's reexamine our random walk for a moment. Clearly the sequence of random movements $w(k_0), \ldots, w(k_0+n-1)$ satisfies the definition of a random process, that is, letting $T = [k_0, k_0+n-1]$, we find that $w(t)$ is a random variable for all $t \in T$. But what about $y(t)$?

To see that $y(t)$ is also a random process, recall that a sum of random variables is also a random variable, that is, for all $k = 0, \ldots, n-1$, the sum $\sum\limits_{i=0}^{k} w(k_0 + i)$ is also a random variable. Furthermore, since the sum of a random variable, such as $\sum\limits_{i=0}^{k} w(k_0+i)$, and a constant, such as $y(k_0)$, is a random variable, $y(t)$ is a random variable for all $t \in T$. Hence $y(t)$ is a random process. □

One further subtlety of a random process is the concept of a *realization.* In the context of our random walk, every path that we might have taken is called a realization. Similarly, if we are testing a noisy sensor in the laboratory, then every possible signal that we might measure as the result of the experiment is called a realization of the random process. More precisely, we have that:

Definition 24.2 Let $X : T \times \Omega \to \mathbb{R}^n$ denote a random process. If ω is fixed, then $X(\cdot, \omega)$ is called a *realization* of the random process X. □

24.2 Properties of Random Processes

Definition 24.3 Let $X : T \times \Omega \to \mathbb{R}^n$ denote a random process, and let $s, t \in T$. The *autocovariance function* of X is the function

$$\gamma(s,t) = \operatorname{cov}\left(X(s), X(t)\right) \tag{24.4}$$

$$\triangleq \mathbb{E}\Big[\Big(X(s) - \mathbb{E}\big[X(s)\big]\Big)\Big(X(t) - \mathbb{E}\big[X(t)\big]\Big)^T\Big] \tag{24.5}$$

□

Definition 24.4 The random process $X : T \times \Omega \to \mathbb{R}^{m \times n}$ is called *stationary* if all of the following hold:

(i) For all $t \in T$: $\quad \texttt{tr}\Big(\mathbb{E}\big[X(t)X^T(t)\big]\Big) < \infty.$

(ii) For all $s, t \in T$: $\quad \mathbb{E}\big[X(s)\big] = \mathbb{E}\big[X(t)\big]$

(iii) For all $s, t, s+k, t+k \in T$: $\quad \gamma(s,t) = \gamma(s+k, t+k)$

□

Definition 24.5 The random process $X : T \times \Omega \to \mathbb{R}^n$ is called *white noise*, or simply *white*, if both of the following hold:

(i) $X(t)$ has the same mean and variance for all $t \in T$.

(ii) For all $s, t \in T$, where $s \neq t$, $X(s)$ and $X(t)$ are uncorrelated.

□

Definition 24.6 The random process $X : T \times \Omega \to \mathbb{R}^n$ is said to be *independent and identically distributed* if both of the following hold:

(i) $X(t)$ has the same probability density function for all $t \in T$.

(ii) For all $s, t \in T$, where $s \neq t$, $X(s)$ and $X(t)$ are independent.

□

Remark 24.5 The term *independent and identically distributed* is commonly abbreviated as *i.i.d.* □

Fact 24.1 Let $X_1, X_2, \ldots$ be a white sequence of Gaussian random variables. Then the sequence $X_1, X_2, \ldots$ is independent and identically distributed.

Proof □

24.3 Convergence of Random Processes

Here we are going to address the convergence of random processes, particularly random sequences. To start, let's revisit the concept of a realization.

Example 24.2 Let $X : \Omega \to \mathbb{R}^n$ denote a random variable which represents the □

A sequence of random variables will always be countably infinite, since they will always be indexed as $X_1, X_2, \ldots$. However, a random process is indexed by time. Hence in the discrete-time case, there is little difference with a sequence of random variables. However, there is obviously a difference in continuous-time.

There is one final point to be made, which particularly pertains system identification... Presumably, you can perform estimation in a continuous-manner, that is, using a continuous-data set. However, nobody does this. Even if you are estimating a continuous-time system, you will have a discrete number of points at which you measured the system. Hence the convergence of your estimates will always be carried out in a discrete setting, that is, we never need to consider the convergence of a random process.

On the other hand, if you are talking about something like the discrete-time Kalman filter, then you DO need to consider truly continuous processes to analyze the behavior of the system. Get ready for the Ito calculus...

Definition 24.7 Let X_N denote a sequence of random variables. Then we say that X_N *converges in distribution* to a random variable X if

$$\lim_{N\to\infty} P\left(X_N \leq x\right) = P\left(X \leq x\right) \tag{24.6}$$

for all of the points $x \in \mathbb{R}$ at which $F(x) = P(X \leq x)$ is continuous. Specifically, we write $X_N \xrightarrow[N\to\infty]{D} X$. □

Definition 24.8 Let X_N denote a sequence of random variables. Then we say that X_N *converges in probability* to a random variable X if

$$\text{for every } \epsilon > 0, \quad \lim_{N\to\infty} P\left(|X_N - X| \leq \epsilon\right) = 0 \tag{24.7}$$

in which case, we write $X_N \xrightarrow[N\to\infty]{P} X$. □

Definition 24.9 Let X_N denote a sequence of random variables. Then we say that X_N *converges with probability one* if

$$P\left(\lim_{N\to\infty} |X_N - X| = 0\right) = 1 \tag{24.8}$$

in which case, we write $X_N \xrightarrow[N\to\infty]{\text{w.p.1}} X$. ❒

Remark 24.6 Convergence with probability one is sometimes called *almost sure* convergence. ❒

Fact 24.2 Let X and $X_1, X_2, \ldots$ be random variables.

(i) If $X_N \xrightarrow[N\to\infty]{\text{w.p.1}} X$, then $X_N \xrightarrow[N\to\infty]{P} X$.

(ii) If $X_N \xrightarrow[N\to\infty]{P} X$, then $X_N \xrightarrow[N\to\infty]{D} X$.

Proof See [12, p. 4] ❒

Fact 24.3 Let X and $X_1, X_2, \ldots$ be random variables, and let g be a continuous function. Then all the following hold:

(i) If $X_N \xrightarrow[N\to\infty]{\text{w.p.1}} X$, then $g(X_N) \xrightarrow[N\to\infty]{\text{w.p.1}} g(X)$.

(ii) If $X_N \xrightarrow[N\to\infty]{P} X$, then $g(X_N) \xrightarrow[N\to\infty]{P} g(X)$.

(iii) If $X_N \xrightarrow[N\to\infty]{D} X$, then $g(X_N) \xrightarrow[N\to\infty]{D} g(X)$.

Proof See [13, pp. 244-246]. ❒

Remark 24.7 Fact 24.3, particularly part (iii), is often referred to as the *continuous mapping theorem.* ❒

Fact 24.4 Let X, Y, $X_1, X_2, \ldots$, and $Y_1, Y_2, \ldots$ be random variables. Then all the following hold:

(i) If $X_N \xrightarrow[N\to\infty]{\text{w.p.1}} X$ and $Y_N \xrightarrow[N\to\infty]{\text{w.p.1}} Y$, then $X_N + Y_N \xrightarrow[N\to\infty]{\text{w.p.1}} X + Y$.

(ii) If $X_N \xrightarrow[N\to\infty]{P} X$ and $Y_N \xrightarrow[N\to\infty]{P} Y$, then $X_N + Y_N \xrightarrow[N\to\infty]{P} X + Y$.

Proof See [13, pp. 247]. □

Fact 24.5 Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables with mean μ, that is, for all $i \geq 1$, $\mathbb{E}[X_i] = \mu$. Then

$$\frac{1}{n}\sum_{k=1}^{n} X_k \xrightarrow[N\to\infty]{P} \mu \quad \text{and} \quad \frac{1}{n}\sum_{k=1}^{n} X_k \xrightarrow[N\to\infty]{\text{w.p.1}} \mu \tag{24.9}$$

Proof The later result is proved in [14, p. 85]. The first result follows from the second. □

24.4 Old Stuff to be checked thoroughly for errors

Fact 24.6 Let X, Y, $X_1, X_2, \ldots$, and $Y_1, Y_2, \ldots$ be random variables. Then all of the following hold:

(i) If $X_N \xrightarrow[N\to\infty]{\text{w.p.1}} X$ and $Y_N \xrightarrow[N\to\infty]{\text{w.p.1}} Y$, then $\left(X_N Y_N^T\right) \xrightarrow[N\to\infty]{\text{w.p.1}} \left(XY^T\right)$.

(ii) If $X_N \xrightarrow[N\to\infty]{\text{w.p.1}} X$ and $z_N \xrightarrow[N\to\infty]{\text{w.p.1}} z$, then $\left(\dfrac{X_N}{z_N}\right) \xrightarrow[N\to\infty]{\text{w.p.1}} \left(\dfrac{X}{z}\right)$

Proof See [15]. □

Fact 24.7 Let $R_1, R_2, \ldots \in \mathbb{R}^{m\times m}$ and $S_1, S_2, \ldots \in \mathbb{R}^{m\times p}$, where

$$R_N \xrightarrow[N\to\infty]{\text{w.p.1}} R, \qquad S_N \xrightarrow[N\to\infty]{\text{w.p.1}} S,$$

and R is nonsingular. Then

$$R_N^{-1} S_N \xrightarrow[N\to\infty]{\text{w.p.1}} R^{-1}S.$$

Proof Let $\alpha \triangleq \det[R]$ and $\alpha_N \triangleq \det[R_N]$. Then since α_N is the finite product and sum of entries of R_N, where each of the entries converges w.p.1, from Fact 24.6 we have that $\alpha_N \xrightarrow[N\to\infty]{\text{w.p.1}} \alpha$. Similarly, letting $T \triangleq \operatorname{adj}[R]$ and $T_N \triangleq \operatorname{adj}[R_N]$, we have that $T_N \xrightarrow[N\to\infty]{\text{w.p.1}} T$. Finally, since R is nonsingular, $\alpha \neq 0$. Hence from Fact 24.6,

$$\frac{1}{\alpha_N} T_N S_N \xrightarrow[N\to\infty]{\text{w.p.1}} \frac{1}{\alpha} TS = R^{-1}S.$$

□

24.5 Quasi-Stationary Signals

Fact 24.8 Let $C \in \mathbb{R}^{p\times p}[\mathbf{q}]$, $D \in \mathbb{R}^{p\times m}[\mathbf{q}]$, and

$$C(\mathbf{q})y(k) = D(\mathbf{q})u(k),$$

where $u \in \mathbb{R}^m$, $y \in \mathbb{R}^p$, $k \geq 1$, and (C, D) is causal and asymptotically stable. Also, let $H \in \mathbb{R}^{p\times m}[\mathbf{q}]$ be the Markov parameter polynomial of (C, D), and let $u = \{u_1, \dots, u_m\} \in \mathbb{R}^m$ and $v = \{v_1, \dots, v_m\} \in \mathbb{R}^m$ be realizations of the independent and identically distributed random processes $\mathcal{U}$ and $\mathcal{V}$, respectively, where $\mathcal{U}$ and $\mathcal{V}$ are mutually independent white processes with bounded second and fourth moments, that is, for all, $i, j, k, \ell \in [1, m]$,

$$\begin{aligned} \mathbb{E}\big[\mathcal{U}_i\mathcal{U}_j\big] &< \infty, & \mathbb{E}\big[\mathcal{V}_i\mathcal{V}_j\big] &< \infty,\\ \mathbb{E}\big[\mathcal{U}_i\mathcal{U}_j\mathcal{U}_k\mathcal{U}_\ell\big] &< \infty, & \mathbb{E}\big[\mathcal{V}_i\mathcal{V}_j\mathcal{V}_k\mathcal{V}_\ell\big] &< \infty. \end{aligned}$$

Then for all $i \in \mathbb{Z}$,

$$\begin{aligned} \frac{1}{N}\sum_{k=1}^{N} y(k+i)y^T(k) &\underset{N\to\infty}{\overset{\text{w.p.1}}{\longrightarrow}} \sum_{j=0}^{\infty} H_{j+i}RH_j,\\ \frac{1}{N}\sum_{k=1}^{N} y(k+i)u^T(k) &\underset{N\to\infty}{\overset{\text{w.p.1}}{\longrightarrow}} H_iR,\\ \frac{1}{N}\sum_{k=1}^{N} y(k+i)v^T(k) &\underset{N\to\infty}{\overset{\text{w.p.1}}{\longrightarrow}} 0_{p\times m},\\ \frac{1}{N}\sum_{k=1}^{N} u(k+i)v^T(k) &\underset{N\to\infty}{\overset{\text{w.p.1}}{\longrightarrow}} 0_{m\times m}, \end{aligned}$$

where $R \in \mathbb{R}^{m\times m}$ is the covariance matrix of $\mathcal{U}$, that is,

$$R \triangleq \mathbb{E}\Big[\mathcal{U}(k)\mathcal{U}^T(k)\Big].$$

Proof See [16, 17]. □

SECTION 25

Rexamination of the Problem

Example 25.1 Consider the model

$$y(k+1) + \alpha_0 y(k) = \beta_1 u(k+1) + \beta_0 u(k) + \gamma_1 w(k+1) + \gamma_0 w(k) \quad (25.1)$$

where $u(k)$ is the sinusoidal signal

$$u(k) = \sin(2\pi f k + \phi) \tag{25.2}$$

and $w(k)$ is the white Gaussian noise process. Then the unknown parameter vector is given by

$$\theta \triangleq \begin{bmatrix} \alpha_0 & \beta_1 & \beta_0 \end{bmatrix} \tag{25.3}$$

where the least-squares estimate $\hat{\theta}$ of θ is given by

$$\begin{aligned}
\hat{\theta} &= \theta + \tilde{\theta}\tilde{\Phi}_{n,N}\Phi_{n,N}^T \left(\Phi_{n,N}\Phi_{n,N}^T\right)^{-1} \\
\tilde{\Phi}_{n,N}\Phi_{n,N}^T &= \sum_{k=1}^{N-1} \begin{bmatrix} w(k+1)y(k) & w(k+1)u(k+1) & w(k+1)u(k) \\ w(k)y(k) & w(k)u(k+1) & w(k)u(k) \end{bmatrix} \\
\Phi_{n,N}\Phi_{n,N}^T &= \sum_{k=1}^{N-1} \begin{bmatrix} y(k)y(k) & y(k)u(k+1) & y(k)u(k) \\ u(k+1)y(k) & u(k+1)u(k+1) & u(k+1)u(k) \\ u(k)y(k) & u(k)u(k+1) & u(k)u(k) \end{bmatrix}
\end{aligned}$$

Then since some signals are quasi-stationary, we have that

$$\begin{aligned}
\frac{\tilde{\Phi}_{n,N}\Phi_{n,N}^T}{N} &\underset{N\to\infty}{\overset{\text{w.p.1}}{\longrightarrow}} \mathbb{E}\begin{bmatrix} W(k+1)Y(k) & W(k+1)U(k+1) & W(k+1)U(k) \\ W(k)Y(k) & W(k)U(k+1) & W(k)U(k) \end{bmatrix} \\
\frac{\Phi_{n,N}\Phi_{n,N}^T}{N} &\underset{N\to\infty}{\overset{\text{w.p.1}}{\longrightarrow}} \mathbb{E}\begin{bmatrix} Y(k)Y(k) & Y(k)U(k+1) & Y(k)U(k) \\ U(k+1)Y(k) & U(k+1)U(k+1) & U(k+1)U(k) \\ U(k)Y(k) & U(k)U(k+1) & U(k)U(k) \end{bmatrix}
\end{aligned}$$

and hence

□

SECTION 26

Measurement and Process Noise

In the ideal case, we would have exact measurements of the inputs u and outputs y of a system, which could be fed directly into a system identification algorithm, as shown in Figure 12. In this case, if we knew how to model the underlying system's dynamics exactly, then we could expect to find a model that exactly fits the measured data.

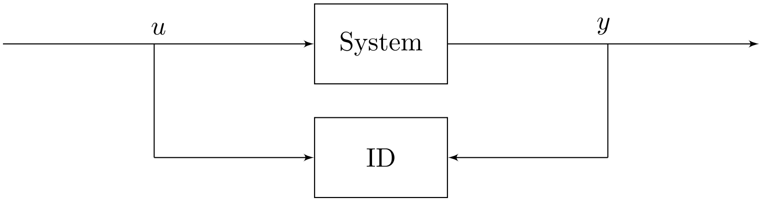


Figure 12: Identification of a system with no noise.

Typically, however, we will have to deal with at least some output measurement noise, also called *sensor noise*, as shown in Figure 13. For instance, if we want to measure the acceleration of a car, that is, $y = \ddot{x}$, then we would use an accelerometer, which will provide us with the noisy measurement $\ddot{x} + v$ of the car's acceleration.

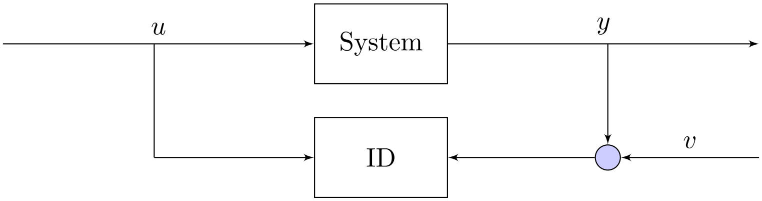


Figure 13: Identification of a system in the presence of the output measurement noise v.

In an even more difficult, but also common scenario, we might have to deal with noise which enters the system's dynamics directly, also called *process noise*, as shown in Figure 14. Whereas sensor noise can be reduced by buying a better sensor, process noise represents actual uncertainty in the system. Typically, process noise is used to represent a physical phenomenon in the system that is not well understood, or would unnecessarily complicate the model. For instance, if we are modelling an airplane's lift, then turbulence might be represented as noise, even though we might be able to model it with a complicated fluid dynamics models.

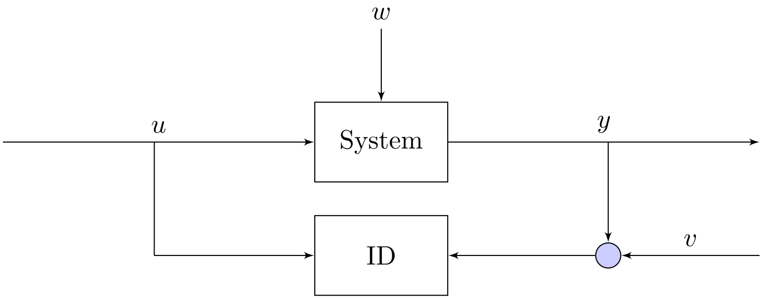


Figure 14: Identification of a system in the presence of the output measurement noise v and process noise w.

Remark 26.1 Figure 14 shows the standard Kalman filtering scenario. □

Finally, in some cases we might need to determine a model for the system between two measured signals, where one signal is viewed as the input, and the other as the output. In this case, we will typically have measurement noise on both the inputs and outputs, as shown in Figure 15. This setup is called *errors-in-variables* (EIV) identification.

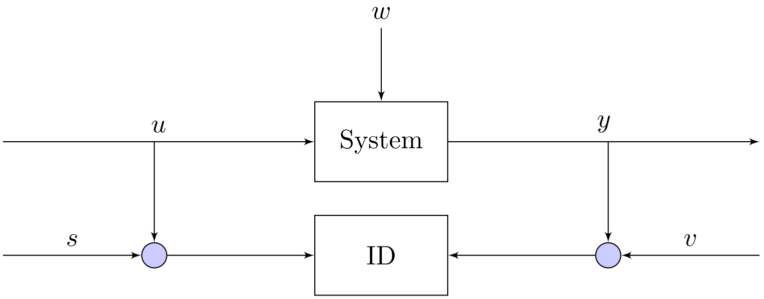


Figure 15: Identification of a system in the presence of the output measurement noise v, process noise w, and input measurement noise s.

Remark 26.2 In this book, the symbol ○ is always used to denote addition (see Figures 13-15). However, in the literature, you might come across cases in which the authors allow for multiplicative noise, in which case the symbol denotes multiplication.

SECTION 27

Noise in State-Space Models

Consider the errors-in-variables identification problem shown in Figure 15, where the system's dynamics are modelled with a time-varying continuous-time model, and the measurements u_m and y_m of u and y have additive noise, that is,

$$\begin{aligned} \dot{x}(t) &= A(t)x(t) + B(t)u(t) + w(t) \\ y(t) &= C(t)x(t) + D(t)u(t) \\ u_m(t) &= u(t) + s(t) \\ y_m(t) &= y(t) + v(t) \end{aligned} \tag{27.1}$$

Then from (27.1), we find that

$$\begin{aligned} \dot{x}(t) &= A(t)x(t) + B(t)u_m(t) + \Big[w(t) - B(t)s(t)\Big] \\ y_m(t) &= C(t)x(t) + D(t)u_m(t) + \Big[v(t) - D(t)s(t)\Big] \end{aligned} \tag{27.2}$$

Remark 27.1 You may be wondering why, instead of (27.2), we don't simply consider models of the form

$$\begin{aligned} \dot{x}(t) &= A(t)x(t) + B(t)u_m(t) + \tilde{w}(t) \\ y_m(t) &= C(t)x(t) + D(t)u_m(t) + \tilde{v}(t) \end{aligned} \tag{27.3}$$

since (27.2) is a special case of (27.3). The reason is that you will find it very difficult to design system identification algorithms for the general system (27.3), where $\tilde{w}$ and u_m can be correlated in an arbitrary manner, and $\tilde{v}$ and y_m can be correlated in an arbitrary manner. It is easier, albeit still quite difficult, to develop algorithms for estimating systems of the form (27.2), where s is correlated with u_m, but w is not. □

SECTION 28

Noise in Polynomial Matrix Models

Consider the errors-in-variables identification problem shown in Figure 15, where the system's dynamics are modelled with a time-varying continuous-time model, and the measurements u_m and y_m of u and y have additive noise, that is,

$$\begin{aligned} A\left(t, \frac{d}{dt}\right) y(t) &= B\left(t, \frac{d}{dt}\right) u(t) + w(t) \\ u_m(t) &= u(t) + s(t) \\ y_m(t) &= y(t) + v(t) \end{aligned} \tag{28.1}$$

Remark 28.1 In the literature you may come across the *ARMAX* models

$$A\left(t, \frac{d}{dt}\right) y(t) = B\left(t, \frac{d}{dt}\right) u(t) + C\left(t, \frac{d}{dt}\right) \tilde{w}(t) \tag{28.2}$$

which is obtained when $w(t) \triangleq C\left(t, d/dt\right) \tilde{w}(t)$ in (28.1). □

PART VI

The Process of System Identification

1) Analyze the measured input and output signals. If there is any obvious distortions that can be easily removed, such as deterministic noise or outliers due to experimental error, remove them.

2) Split the data set into an identification data set, and a validation data set.

3) Assume a model structure for the system's dynamics.

4) Specify how noise enters the system.

5) Make assumptions about the noise process.

6) Select a criterion for fitting the model, that is, construct the cost function that the estimated model should minimize.

7) Estimate the model.

8) Check the cost function residuals to see if there is any structure. For instance, if the residuals contain an obvious sinusoid, then something was not modeled correctly. Change one of the previous steps and start over.

9) Evaluate the cost function of the estimated model on the validation data set. If the estimated model yields poor predictions for the validation data set or if the predictions differ from the validation data (bias, delay, additive sinusoid, etc.) in a structured way, then something was not modeled correctly. Change one of the previous steps and start over.

All systems have a state.

The basic principal of system identification is to assume that we have a system of the form

$$y(t) = \phi(\theta, t, x, u, y, v, w) \tag{28.3}$$

where $\theta \in \mathbb{R}$ represents a vector of parameters that define system. When the system is linear-in the parameters, then

$$y(t) = \theta\phi(t, x, u, y, \theta) \tag{28.4}$$

PART VII

Regression

SECTION 29

Linear Regression

Linear regression is the basis for the simplest forms of system identification. In linear regression, we assume a model structure that is *linear in the parameters*, that is, we assume that, if there is no noise present

$$y(t) = \theta\phi(t, u, y) \tag{29.1}$$

Hence letting τ denote the set of times $t_1, \ldots, t_N$ and letting

$$\begin{align} Y(\tau) &\triangleq \begin{bmatrix} y(t_1) & \cdots & y(t_N) \end{bmatrix} \tag{29.2} \\ \Phi(\tau, u, y) &\triangleq \begin{bmatrix} \phi(t_1, u, y) & \cdots & \phi(t_N, u, y) \end{bmatrix} \tag{29.3} \end{align}$$

we have that

$$Y(\tau) = \theta\Phi(\tau, u, y) \tag{29.4}$$

Therefore, when $\Phi(\tau, u, y)$ has full row rank, we find that

$$\theta = Y(\tau)\Phi^T(\tau, u, y)\Big[\Phi(\tau, u, y)\Phi^T(\tau, u, y)\Big]^{-1} \tag{29.5}$$

29.1 Noise

$$y(t) = \theta\psi(t, u, y) + v(t) \tag{29.6}$$

$$y(t) = \hat{\theta}\phi(t, u, y) + e(t) \tag{29.7}$$

$$\begin{align} Y(\tau) &\triangleq \begin{bmatrix} y(t_1) & \cdots & y(t_N) \end{bmatrix} \tag{29.8} \\ V(\tau) &\triangleq \begin{bmatrix} v(t_1) & \cdots & v(t_N) \end{bmatrix} \tag{29.9} \\ E(\tau) &\triangleq \begin{bmatrix} e(t_1) & \cdots & e(t_N) \end{bmatrix} \tag{29.10} \\ \Psi(\tau, u, y) &\triangleq \begin{bmatrix} \psi(t_1, u, y) & \cdots & \psi(t_N, u, y) \end{bmatrix} \tag{29.11} \\ \Phi(\tau, u, y) &\triangleq \begin{bmatrix} \phi(t_1, u, y) & \cdots & \phi(t_N, u, y) \end{bmatrix} \tag{29.12} \end{align}$$

$$\begin{align} \hat{\theta} &= Y(\tau)\Phi^T(\tau, u, y)\Big[\Phi(\tau, u, y)\Phi^T(\tau, u, y)\Big]^{-1} \tag{29.13} \\ &= \Big[\theta\Psi(\tau, u, y) + V(\tau)\Big]\Phi^T(\tau, u, y)\Big[\Phi(\tau, u, y)\Phi^T(\tau, u, y)\Big]^{-1} \tag{29.14} \end{align}$$

If there is no model error, then

PART VIII

Markov Parameter Identification

SECTION 30

Markov Parameter Identification

These results were taken in vast stretches from the excellent Brockwell and Davis book, *Time Series: Theory and Methods* [16].

Unlike in previous chapters, where we built up to the result, here we are going to present the main result first, and then proceed to work out the conditions which enable the result.

SECTION 31

Problem Statement

Consider the system

$$A(\mathbf{q})y(k) = B(\mathbf{q})u(k), \tag{31.1}$$

where $k \geq 1$, $u \in \mathbb{R}^m$, $y \in \mathbb{R}^p$, $A \in \mathbb{R}^{p\times p}[\mathbf{q}]$ is comonic with $A_0 = I_p$, $B \in \mathbb{R}^{p\times m}[\mathbf{q}]$, and (A, B) is left coprime, causal, and has a degree of n. Furthermore, consider the case where the measurement x of u is corrupted by an additive noise signal v, and the measurement z of y is corrupted by an additive noise signal w, that is,

$$\begin{aligned} x(k) &= u(k) + v(k), \\ z(k) &= y(k) + w(k). \end{aligned} \tag{31.2}$$

as shown in Figure 16.

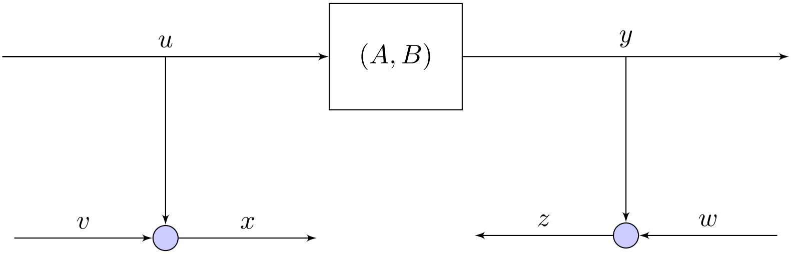


Figure 16: Measurements of the input and output of a linear system in the presence of the measurement noise processes v and w.

Throughout the chapter, we attempt to identify the system (A, B) given only the signals x and z. The identification setup is shown in Figure 17.

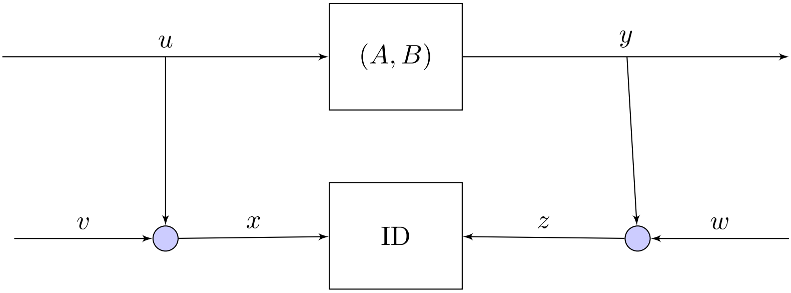


Figure 17: Identification of a linear system in the presence of the measurement noise processes v and w.

SECTION 32

Least-Squares Identification

Here we analyze least-squares estimates in the presence of random noise. We begin by introducing the concept of consistency and semi-consistency as well as the regression notation that we will use henceforth for least-squares identification.

Definition 32.1 For all $N \geq 1$ and $i \in [1, s]$, let $\hat{X}_{i,N} \in \mathbb{R}^{m\times p}$ be an estimate of $X_i \in \mathbb{R}^{m\times p}$. Also, let

$$\hat{\Theta}_N \triangleq \left[\begin{array}{ccc} \hat{X}_{1,N}, & \cdots, & \hat{X}_{s,N} \end{array} \right] \in \mathbb{R}^{m\times ps},$$
$$\Theta \triangleq \left[\begin{array}{ccc} X_1, & \cdots, & X_s \end{array} \right] \in \mathbb{R}^{m\times ps}.$$

Then we say that $\hat{\Theta}_N$ *is a strongly consistent estimate of* Θ if

$$\hat{\Theta}_N \underset{N\to\infty}{\overset{\text{w.p.1}}{\longrightarrow}} \Theta,$$

(see [18]). Furthermore, we say that $\hat{\Theta}_N$ *is a semi-consistent estimate of* Θ if there exists a nonsingular $R \in \mathbb{R}^{p\times p}$ such that

$$\hat{\Theta}_N \underset{N\to\infty}{\overset{\text{w.p.1}}{\longrightarrow}} \Theta \left(I_s \otimes R\right).$$

In this case, for all $i \in [1, s]$,

$$\hat{X}_{i,N} \underset{N\to\infty}{\overset{\text{w.p.1}}{\longrightarrow}} X_i R.$$

□

Remark 32.1 Let $B \in \mathbb{R}^{p\times m}[\mathbf{r}]$, let $R \in \mathbb{R}^{m\times m}$ be nonsingular, and let $\widehat{\theta(B)}_N$ be a semi-consistent estimate of $\theta(B)$, where $B(\mathbf{r})$ has a degree of n and

$$\widehat{\theta(B)}_N \underset{N\to\infty}{\overset{\text{w.p.1}}{\longrightarrow}} \theta(B)\Big(I_{n+1} \otimes R\Big).$$

Then reconstructing $\hat{B}_N(\mathbf{r})$ from its coefficient matrix $\widehat{\theta(B)}_N$, we have that

$$\hat{B}_N(\mathbf{r}) \underset{N\to\infty}{\overset{\text{w.p.1}}{\longrightarrow}} B(\mathbf{r})R.$$

However, while this may not seem useful, note that the zeros of $B(\mathbf{r})$ and the zeros of $B(\mathbf{r})R$ are the same (Appendix **??**). Hence semi-consistent polynomial matrix estimates preserve the zero-structure. □

Notation 32.1 Let $N \geq 1$ and let $x(k) \in \mathbb{R}^m$ for all $k \in [1, N]$. Then for $s \in [1, N]$ and $\mu \in [1, N - s]$, we employ the notation

$$\begin{aligned}
\Psi_{x,s,\mu,N} &\triangleq \begin{bmatrix} x(s+\mu) & \cdots & x(N) \\ \vdots & & \vdots \\ x(s+1) & \cdots & z(N-\mu+1) \end{bmatrix}, \\
\Lambda_{x,s,\mu,N} &\triangleq \begin{bmatrix} x(s) & \cdots & x(N-\mu) \\ \vdots & & \vdots \\ x(1) & \cdots & x(N-s-\mu+1) \end{bmatrix}, \\
\Omega_{x,s,\mu,N} &\triangleq \begin{bmatrix} \Psi_{x,s,\mu,N} \\ \Lambda_{x,s,\mu,N} \end{bmatrix}, \\
\Gamma_{x,s,\mu,N} &\triangleq \begin{bmatrix} x(s+\mu), & \cdots, & x(N) \end{bmatrix}.
\end{aligned}$$

□

Notation 32.2 Let $D \in \mathbb{R}^{p\times m}[\mathbf{q}]$ be given by

$$D(\mathbf{q}) \triangleq D_0 + D_1\mathbf{q} + \cdots + D_s\mathbf{q}^s.$$

Then for $\beta \geq 0$ and $\eta \geq 0$, we employ the notation

$$\begin{aligned}\theta'(D) &\triangleq \left[\begin{array}{ccc} D_1, & \cdots, & D_s \end{array} \right] \in \mathbb{R}^{p \times sm}, \\ \theta_{\beta,\eta}(D) &\triangleq \left[\begin{array}{ccc} D_\beta, & \cdots, & D_{\beta+\eta-1} \end{array} \right] \in \mathbb{R}^{p \times \eta m},\end{aligned}$$

where $D_j = 0_{p \times m}$ for $j > s$. □

32.1 Equation Error Model

A common estimation framework is the equation error approach. Specifically, in the context of the additive measurement noise model (31.2), let the following additional assumption hold:

Assumption 32.1 $v(k) = 0_{m \times 1}$ for all $k \geq 1$. □

Assumption 32.2 There exists a comonic $L \in \mathbb{R}^{p \times p}[\mathbf{r}]$ such that, for all $k \geq 1$,

$$L(\mathbf{q})A(\mathbf{q})w(k) = w'(k),$$

where $w' \in \mathbb{R}^p$ is a realization of the independent and identically distributed, zero-mean, random process $\mathcal{W}'$ with finite covariance. Furthermore, $L_0 = I_p$. □

Then from (31.1)-(31.2) and Assumptions 32.1-32.2, we have the equation error model of (31.1) given by

$$L(\mathbf{q})A(\mathbf{q})z(k) = L(\mathbf{q})B(\mathbf{q})u(k) + w'(k). \tag{32.1}$$

Hence, letting

$$\begin{aligned} C(\mathbf{q}) &\triangleq L(\mathbf{q})A(\mathbf{q}), \\ D(\mathbf{q}) &\triangleq L(\mathbf{q})B(\mathbf{q}), \\ \Theta &\triangleq \left[\begin{array}{cc} \theta(D), & \theta'(C) \end{array} \right], \\ \Phi_N &\triangleq \left[\begin{array}{c} \Omega_{x,s,1,N} \\ -\Lambda_{z,s,1,N} \end{array} \right], \\ Z_N &\triangleq \Gamma_{z,s,1,N}, \\ W'_N &\triangleq \Gamma_{w,s,1,N}, \end{aligned}$$

and letting s denote the degree of (C, D), it follows that

$$Z_N = \Theta\Phi_N + (L_0 A_0)^+ W'_N. \tag{32.2}$$

Next, consider the least-squares estimate $\hat{\Theta}_N$ of Θ given by

$$\hat{\Theta}_N = \underset{\hat{\Theta}_N}{\operatorname{argmin}} \left\| Z_n - \hat{\Theta}_N \Phi_N \right\|_{\mathrm{F}}.$$

Furthermore, let the following assumptions also hold:

Assumption 32.3 For all $k \geq 1$ one of the following holds

i) $u(k)$ is deterministic and bounded.

ii) $u \in \mathbb{R}^m$ is a realization of the random process $\mathcal{U}$, where $\mathcal{U}$ has finite mean and variance.

□

Assumption 32.4 For all $k \geq 1$ and nonnegative i,

$$\mathbb{E}\left[\mathcal{W}'(k+i)\mathcal{U}^T(k) \right] = 0_{p\times m}.$$

□

Assumption 32.5 $(1/N)\Phi_N\Phi_N^T \underset{N\to\infty}{\overset{\text{w.p.1}}{\longrightarrow}} \chi$, where $\chi \in \mathbb{R}^{(m[s+1]+ps)\times(m[s+1]+ps)}$ is nonsingular. □

Assumption 32.6 (C, D) is asymptotically stable. □

Fact 32.1 Let Assumptions 32.1-32.6 hold. Then $\hat{\Theta}_N$ is a strongly consistent estimate of Θ.

Proof First, note that the least-squares estimate $\hat{\Theta}_N$ of Θ satisfies

$$\hat{\Theta}_N \Phi_N \Phi_N^T = Z_N \Phi_N^T.$$

Next, from Assumptions 32.1 and 32.2, we have (32.2) and hence

$$\hat{\Theta}_N \Phi_N \Phi_N^T = \Theta \Phi_N \Phi_N^T + W_N' \Phi_N^T.$$

Furthermore, from Assumptions 32.3-32.6 and Fact 24.8, we have that

$$\begin{aligned} &\frac{1}{N} W_N' \Omega_{x,s,1,N}^T \\ &= \frac{1}{N} \sum_{k=1}^{N-s} \left[\begin{array}{ccc} w'(k+s)x^T(k+s) & \cdots & w'(k+s)x^T(k) \end{array} \right], \\ &\underset{N\to\infty}{\overset{\text{w.p.1}}{\longrightarrow}} 0_{p\times m(s+1)}. \end{aligned}$$

Finally, from (32.1), Assumptions 32.3-32.6, and Fact 24.8, note that

$$\begin{aligned}
&\frac{1}{N} W_N' \Lambda_{z,s,1,N}^T \\
&= \frac{1}{N} \sum_{k=1}^{N-s} \left[\begin{array}{ccc} w'(k+s) z^T(k+s-1) & \cdots & w'(k+s) z^T(k) \end{array} \right], \\
&\xrightarrow[N\to\infty]{\text{w.p.1}} 0_{p\times ps}.
\end{aligned}$$

Hence

$$\frac{1}{N} W_N' \Phi_N^T \xrightarrow[N\to\infty]{\text{w.p.1}} 0_{p\times(sp+m[s+1])},$$

and thus, from Assumption 32.5,

$$\hat{\Theta}_N \xrightarrow[N\to\infty]{\text{w.p.1}} \Theta \Phi_N \Phi_N^T \left(\Phi_N \Phi_N^T \right)^{-1} = \Theta.$$

□

32.2 μ-Markov-Based Least-Squares Estimates of Markov Parameters

Here we show that when the input u and input noise v are white, the Markov parameters can be estimated semi-consistently under fairly general conditions using least-squares and the μ-Markov model.

Consider again the system (31.1) and measurement noise equations (31.2). Also, let

$$\Phi_{s,\mu,N} \triangleq \left[\begin{array}{c} \Omega_{x,s,\mu,N} \\ -\Lambda_{z,s,\mu,N} \end{array} \right], \tag{32.3}$$

$$\tilde{\Phi}_{s,\mu,N} \triangleq \mathbf{q}^{n-s} \left[\begin{array}{c} \Omega_{u,n,\mu,N+n-s} \\ -\Lambda_{y,n,\mu,N+n-s} \end{array} \right], \tag{32.4}$$

$$\Theta_\mu \triangleq \left[\begin{array}{ccc} \theta_{\mu-1}(H), & \theta_{\mu,n}(B^\mu), & \theta_{\mu,n}(A^\mu) \end{array} \right]. \tag{32.5}$$

Next, consider the least-squares estimate

$$\hat{\Theta}_{s,\mu,N} \triangleq \left[\begin{array}{ccc} \widehat{\theta_{\mu-1}(H)}, & \widehat{\theta_{\mu,s}(B^\mu)}, & \widehat{\theta_{\mu,s}(A^\mu)} \end{array} \right],$$

given by

$$\hat{\Theta}_{s,\mu,N} = \underset{\hat{\Theta}_{s,\mu,N}}{\operatorname{argmin}} \left\| \Gamma_{z,s,\mu,N} - \hat{\Theta}_{s,\mu,N} \Phi_{s,\mu,N} \right\|_2^2 . \tag{32.6}$$

Also, consider the following assumptions:

Assumption 32.7 $u \in \mathbb{R}^m$ and $v \in \mathbb{R}^m$ are realizations of the independent and identically distributed processes $\mathcal{U}$ and $\mathcal{V}$, respectively. Furthermore, for all $k \geq 1$, $\mathcal{X}(k) \triangleq \mathcal{U}(k) + \mathcal{V}(k)$. ❒

Assumption 32.8 For all $k \geq 1$ and nonnegative p, the means and covariances of $\mathcal{U}(k)$, $\mathcal{X}(k)$, $\mathcal{U}(k+p)\mathcal{X}^T(k)$, and $\mathcal{X}(k+p)\mathcal{X}^T(k)$ are finite. Furthermore, $R \in \mathbb{R}^{m\times m}$ and $S \in \mathbb{R}^{m\times m}$ given by

$$R \triangleq \mathbb{E}\Big[\mathcal{X}(k)\mathcal{X}^T(k)\Big], \tag{32.7}$$

$$S \triangleq \mathbb{E}\Big[\mathcal{U}(k)\mathcal{X}^T(k)\Big], \tag{32.8}$$

are nonsingular. ❒

Assumption 32.9 $w \in \mathbb{R}^p$ is a realization of the stationary, colored random process $\mathcal{W}$ with finite mean and autocorrelation function, that is, for all $j, k \geq 1$ and nonnegative i,

$$\begin{aligned} \mathbb{E}\Big[\mathcal{W}(k)\Big] &= \mathbb{E}\Big[\mathcal{W}(j)\Big] < \infty, \\ \mathbb{E}\Big[\mathcal{W}(k+i)\mathcal{W}^T(k)\Big] &= \mathbb{E}\Big[\mathcal{W}(j+i)\mathcal{W}^T(j)\Big] < \infty. \end{aligned}$$

❒

Assumption 32.10 The random processes $\mathcal{U}$, $\mathcal{V}$, and $\mathcal{W}$ are independent of each other. ❒

Fact 32.2 Let Assumptions 32.6-32.10 hold. Furthermore, let $s \geq 0$ and $\mu \geq 1$. Then $\widehat{\theta_{\mu-1}(H)}$ is a semi-consistent estimate of $\theta_{\mu-1}(H)$. Specifically,

$$\widehat{\theta_{\mu-1}(H)} \underset{N\to\infty}{\overset{\text{w.p.1}}{\longrightarrow}} \theta_{\mu-1}(H)\Big(I_\mu \otimes SR^{-1}\Big),$$

where R and S are given by (32.7) and (32.8), respectively, and $\widehat{\theta_{\mu-1}(H)}$ is the Markov parameter portion of the least-squares estimate $\hat{\Theta}_{s,\mu,N}$.

Proof First, note that the least-squares estimate $\hat{\Theta}_{s,\mu,N}$ satisfies

$$\begin{aligned}
&\Big(\Gamma_{z,s,\mu,N} - \hat{\Theta}_{s,\mu,N}\Phi_{s,\mu,N}\Big)\Phi^T_{s,\mu,N} \\
&\qquad = \Big(\Gamma_{z,s,\mu,N} - \hat{\Theta}_{s,\mu,N}\Phi_{s,\mu,N}\Big)\begin{bmatrix} \Omega_{x,s,\mu,N} \\ -\Lambda_{z,s,\mu,N} \end{bmatrix}^T \\
&\qquad = \Big(\Gamma_{z,s,\mu,N} - \hat{\Theta}_{s,\mu,N}\Phi_{s,\mu,N}\Big)\begin{bmatrix} \Psi_{x,s,\mu,N} \\ \Lambda_{x,s,\mu,N} \\ -\Lambda_{z,s,\mu,N} \end{bmatrix}^T \\
&\qquad = 0_{p\times(m[\mu+s]+ps)}.
\end{aligned}$$

Next, we examine the subset of previous equations given by

$$\Big(\Gamma_{z,s,\mu,N} - \hat{\Theta}_{s,\mu,N}\Phi_{s,\mu,N}\Big)\Psi^T_{x,s,\mu,N} = 0_{p\times m\mu},$$

where, from (31.1)-(31.2) and (32.3)-(32.5), we have that

$$\Big(\Theta_\mu\tilde{\Phi}_{s,\mu,N} + W_N - \hat{\Theta}_{s,\mu,N}\Phi_{s,\mu,N}\Big)\Psi^T_{x,s,\mu,N} = 0_{p\times m\mu}.$$

Next, from Assumptions 32.6-32.10 and Fact 24.8, we have that

$$\begin{aligned}
&\frac{1}{N}\tilde{\Phi}_{s,\mu,N}\Psi^T_{x,s,\mu,N} \underset{N\to\infty}{\overset{\text{w.p.1}}{\longrightarrow}} \begin{bmatrix} I_\mu\otimes S \\ 0_{n(m+p)\times m\mu} \end{bmatrix}, \\
&\frac{1}{N}W_{s,\mu,N}\Psi^T_{x,s,\mu,N} \underset{N\to\infty}{\overset{\text{w.p.1}}{\longrightarrow}} 0_{p\times m\mu}, \\
&\frac{1}{N}\Phi_{s,\mu,N}\Psi^T_{x,s,\mu,N} \underset{N\to\infty}{\overset{\text{w.p.1}}{\longrightarrow}} \begin{bmatrix} I_\mu\otimes R \\ 0_{s(m+p)\times m\mu} \end{bmatrix}.
\end{aligned}$$

Thus

$$\theta_{\mu-1}(H)\Big(I_\mu\otimes S\Big) - \widehat{\theta_{\mu-1}(H)}\Big(I_\mu\otimes R\Big) \underset{N\to\infty}{\overset{\text{w.p.1}}{\longrightarrow}} 0_{p\times m\mu},$$

and hence from Assumption 32.8, it follows that

$$\widehat{\theta_{\mu-1}(H)} \underset{N\to\infty}{\overset{\text{w.p.1}}{\longrightarrow}} \theta_{\mu-1}(H)\Big(I_\mu\otimes SR^{-1}\Big).$$

□

Corollary 32.1 Let Assumptions 32.1 and 32.6-32.10 hold. Furthermore, let $s\geq 0$ and $\mu\geq 1$. Then $\widehat{\theta_{\mu-1}(H)}$ is a strongly consistent estimate of $\theta_{\mu-1}(H)$, where $\widehat{\theta_{\mu-1}(H)}$ is the Markov parameter portion of the least-squares estimate $\hat{\Theta}_{s,\mu,N}$.

Proof From Assumption 32.1, $v(k) = 0_{m\times 1}$. Thus letting R and S be given by (32.7) and (32.8), respectively, it follows that $R = S$. Finally, from Fact 32.2 it follows that $\widehat{\theta_{\mu-1}(H)} \underset{N\to\infty}{\overset{\text{w.p.1}}{\longrightarrow}} \theta_{\mu-1}(H)$. □

Remark 32.2 In the μ-Markov-based Markov parameter estimates, we need to choose a model degree s as well as the number of Markov parameters μ we would like explicitly displayed in the model (**??**). Furthermore, from Fact 32.2, the degree s of the model does not depend on the degree n of (A, B) in (31.1). In what follows, we refer to s as the model degree which is used in (**??**) and Fact 32.2. □

Example 32.1 Consider the linearized longitudinal model of the T-2 aircraft (**??**) given by

$$\left(1 - 1.862\mathbf{r} + 0.8798\mathbf{r}^2\right) y(k) = \left(-0.009767\mathbf{r} - 0.006026\mathbf{r}^2\right) u(k), \quad (32.9)$$

where $y(2) = y(1) = 0$. Then letting $\mu \triangleq 10$, we have that

$$\begin{aligned}\theta_{\mu-1}(H) &= \begin{bmatrix} H_0, & \cdots, & H_{\mu-1} \end{bmatrix}\\ &= -10^{-2} \times \begin{bmatrix} 0, & 0.977, & 2.42, & 3.65, & 4.66, & 5.47, & 6.09, & 6.52, & 6.79, & 6.90 \end{bmatrix}\end{aligned}$$

Furthermore, let u, v, and w_w be realizations of the independent and identically distributed Gaussian processes $\mathcal{U}$, $\mathcal{V}$, and $\mathcal{W}_w$, respectively, where $\mathcal{U}$, $\mathcal{V}$, and $\mathcal{W}_w$ are all zero mean, independent of each other, and have unit variance. Also, let

$$w(k) = 4w_w(k) + 3w_w(k-1) + 2w_w(k-2) + 1w_w(k-3), \quad (32.10)$$

and consider the additive measurement noise model (31.2). Then from (32.7), $R = 2$, and from (32.8), $S = 1$.

Finally, we let $s = 6$ and estimate $\hat{\Theta}_{s,\mu,N}$ via (32.6) for $M \triangleq 1000$ realizations of the random processes $\mathcal{V}$ and $\mathcal{W}_w$. Then letting $\widehat{\theta^j_{\mu-1}(H)}$ denote the Markov parameter portion of $\hat{\Theta}_{s,\mu,N}$ for each realization $j \in [1, M]$, Figure 18 displays the ensemble average of the estimated Markov parameter error given by

$$\varepsilon \triangleq \frac{1}{\mu M \left\|\theta_{\mu-1}(H)\right\|_2} \sum_{j=1}^{M} \left\| \widehat{\theta^j_{\mu-1}(H)} - \theta_{\mu-1}(H) \right\|_2, \quad (32.11)$$

for several values of N. Specifically, Figure 18 displays two cases: the case where the input measurement noise v is zero, and the case where v is a realization of the zero-mean, unit variance, Gaussian random process $\mathcal{V}$ mentioned previously. In the second case, the estimates are scaled by 2 such that they should in theory be strongly consistent estimates of the true Markov parameters. From Figure 18, it appears that the error ε converges to zero in both both cases, suggesting that $\widehat{\theta_{\mu-1}(H)} \xrightarrow[N\to\infty]{\text{w.p.1}} \theta_{\mu-1}(H)$ when $v = 0$, and $\widehat{\theta_{\mu-1}(H)} \xrightarrow[N\to\infty]{\text{w.p.1}} \theta_{\mu-1}(H)\Big(I_\mu \otimes SR^{-1}\Big) = (1/2)\theta_{\mu-1}(H)$ when v is a realization of the zero-mean, unit variance, Gaussian random process $\mathcal{V}$. Furthermore, from Figure 18, it appears that the estimates converge more slowly when $v \neq 0$. ◻

Figure 18: Comparison of the ensemble average of the error in the estimated Markov parameters of (32.9) when y is measured with the additive colored noise signal (32.10), and either $v(k) = 0$ or v a realization of a zero-mean, unit variance, Gaussian random variable, where $s = 6$ and $\mu = 10$.

SECTION 33

Recovering the Linear System from Markov Parameters

Here we show how to obtain semi-consistent system estimates from semi-consistent Markov parameter estimates.

Fact 33.1 Consider the system (31.1), where $H \in \mathbb{R}^{p\times m}[\mathbf{r}]$ is the Markov parameter polynomial of (A, B). Also, let $n^\star$ be the degree of the quasi-scalar multiple of (A, B) given by Proposition **??**, and let $\bar{n} \geq n^\star$, let $R \in \mathbb{R}^{m\times m}$ be nonsingular, and let $\hat{H}_{0,N}, \ldots, \hat{H}_{2\bar{n},N} \in \mathbb{R}^{p\times m}$ be semi-consistent estimates of $H_0, \ldots, H_{2\bar{n}}$, respectively. Specifically, for all $i \in [0, 2\bar{n}]$, let

$$\hat{H}_{i,N} \xrightarrow[N\to\infty]{\text{w.p.1}} H_i R.$$

Finally, let

$$\begin{aligned}\hat{H}_N(\mathbf{q}) &\triangleq \sum_{i=0}^{2s} \hat{H}_{i,N}\mathbf{q}^i,\\ \hat{\theta}_{C,N} &\triangleq \left[\begin{array}{ccc} \hat{C}_{1,N}, & \cdots, & \hat{C}_{\bar{n},N} \end{array}\right],\\ \hat{\theta}_{D,N} &\triangleq \left[\begin{array}{ccc} \hat{D}_{0,N}, & \cdots, & \hat{D}_{\bar{n},N} \end{array}\right],\end{aligned}$$

where $\hat{\theta}_{C,N}$ and $\hat{\theta}_{D,N}$ minimize

$$J_N = \left\| \theta_{2\bar{n}}(\hat{H}_N) - \left[\begin{array}{cc} \hat{\theta}_{C,N}, & \hat{\theta}_{D,N} \end{array}\right] \overline{\mathcal{K}}_{\bar{n},\bar{n}}\big(\hat{H}_N\big) \right\|_{\mathrm{F}}, \tag{33.1}$$

and where $\overline{\mathcal{K}}_{\bar{n},\bar{n}}\big(\hat{H}_N\big)$ is given by Definition **??**. If $\hat{C}_N \in \mathbb{R}^{p\times p}[\mathbf{r}]$ and $\hat{D}_N \in \mathbb{R}^{p\times m}[\mathbf{r}]$ are given by

$$\begin{aligned}\hat{C}_N(\mathbf{q}) &\triangleq I_p + \hat{C}_{1,N}\mathbf{q} + \cdots + \hat{C}_{\bar{n},N}\mathbf{q}^{\bar{n}},\\ \hat{D}_N(\mathbf{q}) &\triangleq \hat{D}_{0,N} + \hat{D}_{1,N}\mathbf{q} + \cdots + \hat{D}_{\bar{n},N}\mathbf{q}^{\bar{n}},\end{aligned}$$

the $\left(\hat{C}_N, \hat{D}_N\right)$ converges with probability one to a multiple of (A, BR).

Proof Since R is nonsingular and $\hat{H}_{i,N} \underset{N\to\infty}{\overset{\text{w.p.1}}{\longrightarrow}} H_i R$ for all $i \in [0, 2\bar{n}]$,

$$\overline{\mathcal{K}}_{\bar{n},\bar{n}}\big(\hat{H}_N\big) \underset{N\to\infty}{\overset{\text{w.p.1}}{\longrightarrow}} \overline{\mathcal{K}}_{\bar{n},\bar{n}}\big(HR\big).$$

Hence

$$\left[\begin{array}{cc} \theta\left(\hat{C}_N\right), & \theta\left(\hat{D}_N\right) \end{array}\right] \overline{\mathcal{K}}_{\bar{n},\bar{n}}\big(\hat{H}_N\big) \underset{N\to\infty}{\overset{\text{w.p.1}}{\longrightarrow}} 0_{p\times m(2\bar{n}+1)},$$

and therefore, from Proposition **??**,

$$\hat{C}_N(\mathbf{r})H(\mathbf{r})R - \hat{D}_N(\mathbf{r}) \underset{N\to\infty}{\overset{\text{w.p.1}}{\longrightarrow}} 0_{p\times m}.$$

Finally, since $\hat{C}_N(\mathbf{r})$ is comonic, $C(\mathbf{r})$ has full normal rank, and thus from Theorem **??**, $(\hat{C}_N, \hat{D}_N)$ converges with probability one to a multiple of (A, BR). □

Remark 33.1 There may exist multiple minimizers $\hat{\theta}_{C,N}$ and $\hat{\theta}_{D,N}$ of (33.1), even if $\bar{n} = n$. This is due to the fact that in MIMO polynomial matrix models, there may exist more than one parameterization of the same system. However, every solution will still be a multiple of (A, BR). □

Remark 33.2 Note that if R is nonsingular, then the zeros of $B(\mathbf{r})R$ are the same as the zeros of $B(\mathbf{r})$ (see Appendix **??**). □

Example 33.1 Consider the linearized longitudinal model of the T-2 aircraft (**??**) in Example 32.1, where

$$\left(1 - 1.862\mathbf{r} + 0.8798\mathbf{r}^2\right) y(k) = \left(-0.009767\mathbf{r} - 0.006026\mathbf{r}^2\right) u(k).$$

Furthermore, let $y(2) = y(1) = 0$, and let u, v, and w_w be realizations of the independent and identically distributed Gaussian processes $\mathcal{U}$, $\mathcal{V}$, and $\mathcal{W}_w$, respectively, where $\mathcal{U}$, $\mathcal{V}$, and $\mathcal{W}_w$ are all zero mean and independent of each other. However, now let the variances of $\mathcal{U}$, $\mathcal{V}$, and $\mathcal{W}_w$ be 1, 1/20, and 1/10000, respectively. Also, let

$$w(k) = 4w_w(k) + 3w_w(k-1) + 2w_w(k-2) + 1w_w(k-3),$$

and consider the additive measurement noise model (31.2). Then the signal to noise ratios of x and z are both approximately 5.

Finally, let $s = 6$, $\mu = 10$, and let $\widehat{\theta_{\mu-1}(H)}$ be the Markov parameter portion of the least-squares estimate $\hat{\Theta}_{s,\mu,N}$ given by (32.6). Furthermore, let $\bar{n} = 3$ and let the Markov parameters estimates in $\widehat{\theta_{\mu-1}(H)}$ be used to estimate $\hat{C}_N(\mathbf{r})$ and $\hat{D}_N(\mathbf{r})$ in Fact 33.1. Then Figure 19 displays the estimates of the coefficients of $\hat{C}_N(\mathbf{r})$ given by Fact 33.1 along with their limiting values

$$\hat{C}_N(\mathbf{r}) \underset{N\to\infty}{\overset{\text{w.p.1}}{\longrightarrow}} 1 - 1.862\mathbf{r} + 0.8798\mathbf{r}^2,$$

for $N = 10^2, \ldots, 10^4$ and a sample realization of $\mathcal{U}$, $\mathcal{V}$, and $\mathcal{W}_w$. Figure 20 displays the estimates the coefficients of $\hat{D}_N(\mathbf{r})$ given by Fact 33.1 along with their limiting values

$$\hat{D}_N(\mathbf{r}) \underset{N\to\infty}{\overset{\text{w.p.1}}{\longrightarrow}} \left[\frac{20}{21}\right] \left(-0.009767\mathbf{r} - 0.006026\mathbf{r}^2\right),$$

for $N = 10^2, \ldots, 10^4$ and the same realization of $\mathcal{U}$, $\mathcal{V}$, and $\mathcal{W}_w$. Note that the scaling 20/21 in the coefficients of $\hat{D}_N(\mathbf{r})$ reflects the fact that the input variance is 1/20. □

Figure 19: Comparison of the coefficients of $\hat{C}_N(\mathbf{r})$ (solid line) along with their limiting values (dashed line) as N increases, where $\hat{C}_N(\mathbf{r})$ is given by Fact 33.1.

Figure 20: Comparison of the coefficients of $\hat{D}_N(\mathbf{r})$ (solid line) along with their limiting values (dashed line) as N increases, where $\hat{D}_N(\mathbf{r})$ is given by Fact 33.1.

SECTION 34

Conclusions

We addressed the issue of identification in the presence of random noise. Specifically, we studied the consistency of the estimates in two scenarios, namely, the equation-error framework and the case where the input and input noise were white. In the latter case, we presented an approach based on using least-squares with a μ-Markov model. Finally, we introduced the concept of semi-consistency and showed how, using the techniques developed in Chapter **??**, one could obtain semi-consistent linear system estimates from semi-consistent Markov parameter estimates.

PART IX

Frequency Domain Identification

SECTION 35

The Fourier Transform

PART X

Realizations of Various Models

Theorem 35.1 If $\boldsymbol{\rho}$ is either the differentiation operator $\mathbf{d/dt}$ or the backward-shift operator $\boldsymbol{\rho}$, then the polynomial matrix model (**??**) is linear.

Proof We will show that the polynomial matrix models are equivalent to state-space models □

Remark 35.1 Theorem 35.1 only shows that polynomial matrix models are linear for the differentiation operator and the back-shift operator since we can directly link these to the state-space models. To show that polynomial matrix models are linear for an arbitrary operator $\boldsymbol{\rho}$, we would have to first establish that the solution y of (**??**) is unique for all inputs and initial conditions. □

There are two questions that naturally arise

1. Why not call them just *polynomial models*, or *transfer function models*?

2. What is the *state* of a polynomial matrix model?

Furthermore, finite impulse response models form a significant subset of discrete-time models in which the output will equal exactly zero after a finite number of time steps.

Transfer functions and state-space models provide alternative and complementary parametric representations for multivariable linear systems, with transfer function models providing an easy-to-work-with link between the two [1316]. Similarly, frequency response models and Markov parameter models provide additional, albeit nonparametric, representations for the same systems [2, 23].

SECTION 36

Connection with State-Space Models

Here we consider the connection between polynomial matrix models, state-space models, and Markov parameters. Specifically, we review the well-known method of obtaining a polynomial matrix model from a state-space model, and then show that, using the Markov parameters of the state-space model, we can obtain the same polynomial matrix model using the algorithms in the present paper, particularly Proposition **??**. Furthermore, we show that all of the same rank properties presented in Proposition **??** still hold when the Markov parameters are generated from a state-space model, where $n^\star$ is replaced by the order of the state-space model which generates the Markov parameters.

Theorem 36.1 Consider the state-space system

$$\begin{aligned} x(t) &= \boldsymbol{\rho}\tilde{A}x(t) + \boldsymbol{\rho}\tilde{B}u(t), \\ y(t) &= \tilde{C}x(t) + \tilde{D}u(t), \end{aligned}$$

where $\tilde{A} \in \mathbb{R}^{n\times n}$, $\tilde{B} \in \mathbb{R}^{n\times m}$, $\tilde{C} \in \mathbb{R}^{p\times n}$, and $\tilde{D} \in \mathbb{R}^{p\times m}$, $x \in \mathbb{R}^n$ is the state, $u \in \mathbb{R}^m$ is the input, and $y \in \mathbb{R}^p$ is the output. Also, let

$$\begin{aligned} A(\boldsymbol{\rho}) &\triangleq det\left[I_n - \boldsymbol{\rho}\tilde{A}\right], \\ E(\boldsymbol{\rho}) &\triangleq adj\left[I_n - \boldsymbol{\rho}\tilde{A}\right], \\ B(\boldsymbol{\rho}) &\triangleq \boldsymbol{\rho}\tilde{C}E(\boldsymbol{\rho})\tilde{B} + A(\boldsymbol{\rho})\tilde{D}. \end{aligned}$$

Then $A(\boldsymbol{\rho})y(t) = B(\boldsymbol{\rho})u(t)$.

Proof

$$A(\boldsymbol{\rho})y(t) = \tilde{C}A(\boldsymbol{\rho})x(t) + A(\boldsymbol{\rho})\tilde{D}u(t) = \tilde{C}\left[E(\boldsymbol{\rho})\boldsymbol{\rho}\tilde{B}u(t)\right] + A(\boldsymbol{\rho})\tilde{D}u(t) = B(\boldsymbol{\rho})u(t).$$

□

Definition 36.1 Let $\tilde{A} \in \mathbb{R}^{n\times n}$, $\tilde{B} \in \mathbb{R}^{n\times m}$, $\tilde{C} \in \mathbb{R}^{p\times n}$, and $\tilde{D} \in \mathbb{R}^{p\times m}$. Also, for $i \geq 1$, let

$$H_0 \triangleq \tilde{D}, \quad H_1 \triangleq \tilde{C}\tilde{B}, \quad H_2 = \tilde{C}\tilde{A}\tilde{B}, \quad \cdots, \quad H_i \triangleq \tilde{C}\tilde{A}^{i-1}\tilde{B}.$$

Then H_j is the j^{th} *Markov parameter* of $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$, and

$$H(\boldsymbol{\rho}) \triangleq \sum_{j=0}^{\infty} H_j \mathbf{r}^j,$$

is the *Markov parameter polynomial* of $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$. □

Theorem 36.2 Consider the controllable state-space model

$$\begin{aligned} x(t) &= \boldsymbol{\rho}\tilde{A}x(t) + \boldsymbol{\rho}\tilde{B}u(t), \\ y(t) &= \tilde{C}x(t) + \tilde{D}u(t), \end{aligned}$$

where $\tilde{A} \in \mathbb{R}^{n\times n}$, $\tilde{B} \in \mathbb{R}^{n\times m}$, $\tilde{C} \in \mathbb{R}^{p\times n}$, and $\tilde{D} \in \mathbb{R}^{p\times m}$, $x \in \mathbb{R}^n$ is the state, $u \in \mathbb{R}^m$ is the input, and $y \in \mathbb{R}^p$ is the output. Furthermore, let $\bar{n} \geq n$ and let $H \in \mathbb{R}_{\infty}^{p\times m}[\boldsymbol{\rho}]$ be the Markov parameter polynomial of $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$. Then for all nonnegative t,

$$rank\Big[\mathcal{K}_{t,n}(H)\Big] = rank\Big[\mathcal{K}_{t,\bar{n}}(H)\Big], \tag{36.1}$$

$$rank\Big[\overline{\mathcal{K}}_{t,n}(H)\Big] = rank\Big[\overline{\mathcal{K}}_{t,\bar{n}}(H)\Big]. \tag{36.2}$$

Furthermore, letting

$$\begin{aligned} A(\boldsymbol{\rho}) &\triangleq det\left[I_n - \boldsymbol{\rho}\tilde{A}\right], \\ E(\boldsymbol{\rho}) &\triangleq adj\left[I_n - \boldsymbol{\rho}\tilde{A}\right], \\ B(\boldsymbol{\rho}) &\triangleq \boldsymbol{\rho}\tilde{C}E(\boldsymbol{\rho})\tilde{B} + A(\boldsymbol{\rho})\tilde{D}, \end{aligned}$$

then

$$\left[\begin{array}{cc} \theta_n\left(AI_p\right) & -\theta_n\left(B\right) \end{array}\right] \mathcal{K}_{n,\bar{n}}\left(H\right) = 0_{p\times m(n+\bar{n}+1)}, \tag{36.3}$$

and there exists a nonnegative $s \leq n$ such that

$$rank\Big[\overline{\mathcal{K}}_{s,n}(H)\Big] = rank\Big[\mathcal{K}_{s,n}(H)\Big]. \tag{36.4}$$

Proof First, note that from Definition **??** and Definition 36.1, for all $\bar{n} \geq n$ and $t \geq 0$, we have that

$$\begin{aligned} \mathcal{K}_{t,\bar{n}}(H) &= \left[\begin{array}{cc} \mathcal{T}_t(H) & \mathcal{O}_t\left(\tilde{A},\tilde{C}\right)\mathcal{C}_{\bar{n}}\left(\tilde{A},\tilde{B}\right) \\ I_{m(t+1)} & 0_{m(t+1)\times m\bar{n}} \end{array}\right], \\ \mathcal{O}_t\left(\tilde{A},\tilde{C}\right) &\triangleq \left[\begin{array}{cccc} \left(\tilde{C}\tilde{A}^t\right)^T & \cdots & \left(\tilde{C}\tilde{A}\right)^T & \tilde{C}^T \end{array}\right]^T, \\ \mathcal{C}_{\bar{n}}\left(\tilde{A},\tilde{B}\right) &\triangleq \left[\begin{array}{cccc} \tilde{B} & \tilde{A}\tilde{B} & \cdots & \tilde{A}^{\bar{n}-1}\tilde{B} \end{array}\right], \end{aligned}$$

where $\mathcal{O}_n\left(\tilde{A},\tilde{C}\right)$ is the reordered observability matrix of $(\tilde{A},\tilde{C})$, and $\mathcal{C}_n\left(\tilde{A},\tilde{B}\right)$ is the controllability matrix of $(\tilde{A},\tilde{B})$. Furthermore, since $(\tilde{A},\tilde{B},\tilde{C},\tilde{D})$ is controllable, then for all $\bar{n} \geq n$, $\mathcal{C}_{\bar{n}}\left(\tilde{A},\tilde{B}\right)$ has full row rank. Hence for all $\bar{n} \geq n$, it follows that

$$rank\left[\mathcal{O}_t\left(\tilde{A},\tilde{C}\right)\mathcal{C}_n\left(\tilde{A},\tilde{B}\right)\right] = rank\left[\mathcal{O}_t\left(\tilde{A},\tilde{C}\right)\mathcal{C}_{\bar{n}}\left(\tilde{A},\tilde{B}\right)\right] = rank\left[\mathcal{O}_t\left(\tilde{A},\tilde{C}\right)\right] \leq n$$

that is, the final $m(\bar{n}-n)$ columns of $\mathcal{K}_{t,\bar{n}}(H)$ are in the column space of the previous mn columns and therefore (36.1). Similarly, we have (36.2).

Next, note that

$$B(\boldsymbol{\rho}) - A(\boldsymbol{\rho})H(\boldsymbol{\rho}) = \boldsymbol{\rho}\tilde{C}E(\boldsymbol{\rho})\tilde{B} + A(\boldsymbol{\rho})\tilde{D} - A(\boldsymbol{\rho})\left(\tilde{D} + \sum_{i=1}^{\infty}\tilde{C}\tilde{A}^{i-1}\tilde{B}\boldsymbol{\rho}^i\right) = \boldsymbol{\rho}\tilde{C}\left(E(\boldsymbol{\rho}) -\right.$$

Furthermore, since

$$\left[I_n - \boldsymbol{\rho}\tilde{A}\right]\sum_{i=0}^{\infty}\tilde{A}^i\boldsymbol{\rho}^i = I_n,$$

it follows that

$$\left[I_n - \boldsymbol{\rho}\tilde{A}\right]\left(E(\boldsymbol{\rho}) - A(\boldsymbol{\rho})\sum_{i=0}^{\infty}\tilde{A}^i\boldsymbol{\rho}^i\right) = \mathtt{det}\left[I_n - \boldsymbol{\rho}\tilde{A}\right] - A(\boldsymbol{\rho}) = 0_{n\times n},$$

where, since $\left[I_n - \boldsymbol{\rho}\tilde{A}\right]$ is regular, from Fact **??**, $\left[I_n - \boldsymbol{\rho}\tilde{A}\right]$ has full row rank. Hence, from Fact **??**,

$$E(\boldsymbol{\rho}) - A(\boldsymbol{\rho})\sum_{i=0}^{\infty}\tilde{A}^i\boldsymbol{\rho}^i = 0_{n\times n},$$

and therefore

$$B(\boldsymbol{\rho}) - A(\boldsymbol{\rho})H(\boldsymbol{\rho}) = \boldsymbol{\rho}\tilde{C}\left(0_{n\times n}\right)\tilde{B} = 0_{p\times m},$$

that is, $A(\boldsymbol{\rho})H(\boldsymbol{\rho}) = B(\boldsymbol{\rho})$.

Finally, note that $A(\boldsymbol{\rho})$ has degree less than or equal n from the definition of the determinant, and from the definition of the adjugate in terms of the cofactor matrix, it follows that $E(\boldsymbol{\rho})$ has degree less than or equal $n-1$. Hence $B(\boldsymbol{\rho})$ has degree less than or equal to n. Therefore, since (A, B) has degree less than or equal n, and $A(\boldsymbol{\rho})H(\boldsymbol{\rho}) = B(\boldsymbol{\rho})$, we have (36.3). Furthermore, since

$$A(0) = \mathtt{det}\left[I_n - 0\times\tilde{A}\right] = 1 = A_0,$$

we have (36.4). □

PART XI

Kalman Filtering

SECTION 37

Continuous-Time Kalman Filter

Consider the time-varying, continuous-time state-space system

$$\begin{aligned} \dot{x}(t) &= A(t)x(t) + B(t)u(t) + w(t) \\ y(t) &= C(t)x(t) + D(t)u(t) + v(t) \end{aligned} \tag{37.1}$$

where

- $x \in \mathbb{R}^n$ denotes an unknown state vector.
- $u \in \mathbb{R}^m$ denotes a known (or measured) input.
- $y \in \mathbb{R}^p$ denotes a known (or measured) output.
- $w \in \mathbb{R}^n$ denotes an unknown process noise vector.
- $v \in \mathbb{R}^p$ denotes an unknown measurement noise vector.
- $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$, and $D \in \mathbb{R}^{p\times m}$ are known.

Then the purpose of a continuous-time Kalman filter is to produce an estimate $\hat{x}(t)$ of $x(t)$ given the signals u and y, the system matrices (A, B, C, D), some knowledge about the noise statistics, and an initial state estimate $\hat{x}(t_0)$. Specifically, the Kalman filter estimate $\hat{x}(t)$ of $x(t)$ is given by

$$\dot{\hat{x}}(t) = A(t)\hat{x}(t) + B(t)u(t) + G(t)\Big[y(t) - C(t)\hat{x}(t) - D(t)u(t)\Big] \tag{37.2}$$

where the gain $G \in \mathbb{R}^{p\times n}$ is chosen to minimize the change in the expected value of the estimation error, that is,

$$J(t) = \frac{d}{dt}\left[\mathbb{E}\Big\|\hat{x}(t) - x(t)\Big\|_2^2\right] \tag{37.3}$$

Remark 37.1 The order of differentiation and expectation in (37.3) is very important. In general,

$$\frac{d}{dt}\left[\mathbb{E}\Big\|\hat{x}(t) - x(t)\Big\|_2^2\right] \neq \mathbb{E}\left[\frac{d}{dt}\Big\|\hat{x}(t) - x(t)\Big\|_2^2\right] \tag{37.4}$$

Remark 37.2 The gain $G(t)$ which minimizes (37.3) is called the *Kalman gain.* It ensures that the change in the expected magnitude of $\hat{x}(t) - x(t)$ decreases as much as possible, that is, it attempts to move the estimate $\hat{x}(t)$ closer to the true state $x(t)$. The caveat is that we cannot always ensure that the expected magnitude of $\hat{x}(t) - x(t)$ is actually decreasing. For instance, if $J(t) > 0$ for all $G(t)$, then we expect the estimate to get worse, even if we use the Kalman gain. In this case, the Kalman gain is *the best of the worst*, that is, although the Kalman filter estimates will degrade over time, they are better than if we had chosen a different gain $G(t)$ in (37.2). □

To compute the Kalman gain, let $e(t)$ denote the instantaneous estimation error, that is, let

$$e(t) \triangleq \hat{x}(t) - x(t) \tag{37.5}$$

Then the Kalman filter cost function is given by

$$J(t) = \frac{d}{dt}\left[\mathbb{E}\left[e^T(t)e(t)\right]\right] = \frac{d}{dt}\texttt{tr}\left(\mathbb{E}\left[e(t)e^T(t)\right]\right) \tag{37.6}$$

where $\texttt{tr}(\cdot)$ denotes the matrix trace (see Appendix E). Hence, letting

$$P(t) \triangleq \mathbb{E}\left[e(t)e^T(t)\right] \tag{37.7}$$

we have that the Kalman filter cost function is given by

$$J(t) = \texttt{tr}\left(\dot{P}(t)\right) \tag{37.8}$$

Next, from (37.8), note that we are going to need an expression for $\dot{P}(t)$ in order to calculate the Kalman gain. We accomplish this using the following procedure:

(1) Compute the error dynamics $\dot{e}(t)$ from (37.1) and (37.2).

(2) Use the variation of parameters formula (2.2) from Section 2 to determine an equation for $e(t)$.

(3) Make some assumptions about the noise processes and compute $P(t)$.

(4) Differentiate $P(t)$.

We begin with the error dynamics. Specifically, from (37.1) and (37.2), note that

$$\dot{e}(t) = \Big[A(t) - G(t)C(t)\Big]e(t) + G(t)v(t) - w(t) \tag{37.9}$$

Hence the current error $e(t)$ is given in terms of the initial error $e(t_0)$ by

$$e(t) = \Phi(t, t_0)e(t_0) + \int_{t_0}^{t} \Phi(t, \tau)\Big[G(\tau)v(\tau) - w(\tau)\Big]d\tau \tag{37.10}$$

where the state transition matrix $\Phi(t, t_0)$ is the solution of

$$\dot{\Phi}(t, t_0) = \Big[A(t) - G(t)C(t)\Big]\Phi(t, t_0) \tag{37.11}$$

subject to the initial condition $\Phi(t_0, t_0) = I$.

Next, note that the covariance matrix $P(t)$ is given by

$$\begin{aligned} P(t) = \mathbb{E}\Bigg[&\left(\Phi(t, t_0)e(t_0) + \int_{t_0}^{t} \Phi(t, \tau)\Big[G(\tau)v(\tau) - w(\tau)\Big]d\tau\right) \\ &\left(\Phi(t, t_0)e(t_0) + \int_{t_0}^{t} \Phi(t, \phi)\Big[G(\phi)v(\phi) - w(\phi)\Big]d\phi\right)^T\Bigg] \end{aligned} \tag{37.12}$$

where the state-transition matrices $\Phi(t, t_0)$, $\Phi(t, \tau)$, and $\Phi(t, \phi)$ must remain inside the expectation operator since they might be correlated with the noise processes and initial estimation error $e(t_0)$. Hence to simplify the covariance equation, we introduce the assumption:

Assumption 37.1 For all $t \in \mathbb{R}$, $A(t)$, $C(t)$, and $G(t)$ are deterministic. □

From Assumption 37.1, it follows that the state-transition matrix $\Phi(t, t_0)$ is deterministic since the differential equation (37.11) only involves deterministic matrices. Therefore, using Assumption 37.1, we find that

$$P(t) = \Phi(t, t_0)P(t_0)\Phi^T(t, t_0) + \tilde{P}(t) + \tilde{P}^T(t) + \tilde{\tilde{P}}(t) \tag{37.13}$$

where $\tilde{\tilde{P}}(t)$ contains the product of the two integral terms, and

$$\tilde{P}(t) \triangleq \left(\int_{t_0}^{t} \Phi(t, \tau)\mathbb{E}\Big[G(\tau)v(\tau)e^T(t_0) - w(\tau)e^T(t_0)\Big]d\tau\right)\Phi^T(t, t_0) \tag{37.14}$$

Remark 37.3 Technically, $P(t)$ should only be called the covariance matrix if $\mathbb{E}[e(t)] = 0$ for all $t \in \mathbb{R}$. From (37.10) and Assumption 37.1, we see that this is indeed the case if, for all $t \in \mathbb{R}$:

$$\mathbb{E}[e(t_0)] = 0, \quad \mathbb{E}[v(t)] = 0, \quad \mathbb{E}[w(t)] = 0 \tag{37.15}$$

However, since these assumptions are not required for the rest of the development, we do not explicitly require them. Just be aware that without (37.15), $P(t)$ is not technically the *covariance* matrix. □

At this point, we clearly need to make some assumptions about the terms $v(t)e^T(t_0)$ and $w(t)e^T(t_0)$ to be able to evaluate $\tilde{P}(t)$. The standard assumption is that the noise processes are uncorrelated with the initial estimation error, that is:

Assumption 37.2 For all $t \in \mathbb{R}$,

$$\mathbb{E}\Big[w(t)e^T(t_0)\Big] = 0, \qquad \mathbb{E}\Big[v(t)e^T(t_0)\Big] = 0 \tag{37.16}$$

□

Specifically, using Assumptions 37.1-37.2, we find that $\tilde{P}(t) = 0$ and

$$P(t) = \Phi(t, t_0)P(t_0)\Phi^T(t, t_0) + \tilde{\tilde{P}}(t) \tag{37.17}$$

where $\tilde{\tilde{P}}(t)$ is given by

$$\tilde{\tilde{P}}(t) \triangleq \int_{t_0}^{t}\int_{t_0}^{t} \Phi(t, \tau)Z(\tau, \phi)\Phi^T(t, \phi)d\tau d\phi \tag{37.18}$$

$$Z(\tau, \phi) \triangleq \mathbb{E}\left[\Big[G(\tau)v(\tau) - w(\tau)\Big]\Big[G(\phi)v(\phi) - w(\phi)\Big]^T\right] \tag{37.19}$$

The final assumptions we need to make are about the terms $v(\tau)v^T(\phi)$, $w(\tau)w^T(\phi)$, and $v(\tau)w^T(\phi)$. In the standard Kalman filter, one assumes that w and v are stationary white noise processes that are uncorrelated with each other, that is:

Assumption 37.3 There exists a positive semi-definite $P_w \in \mathbb{R}^{n\times n}$ and $P_v \in \mathbb{R}^{p\times p}$, such that for all $t, \tau \in \mathbb{R}$:

$$\begin{aligned}\mathbb{E}\Big[\ w(t)w^T(\tau)\ \Big] &= \delta(t-\tau)P_w\\ \mathbb{E}\Big[\ v(t)v^T(\tau)\ \Big] &= \delta(t-\tau)P_v\\ \mathbb{E}\Big[\ v(t)w^T(\tau)\ \Big] &= 0\end{aligned} \tag{37.20}$$

where δ denotes the Dirac delta function. □

Hence, using Assumptions 37.1-37.3, we find that

$$\tilde{\tilde{P}}(t) = \int_{t_0}^{t}\int_{t_0}^{t} \Phi(t,\tau)\Big[\delta(\tau-\phi)G(\tau)P_vG^T(\phi) + \delta(\tau-\phi)P_w\Big]\Phi^T(t,\phi)d\tau d\phi$$

and therefore

$$P(t) = \Phi(t,t_0)P(t_0)\Phi^T(t,t_0) + \int_{t_0}^{t}\Phi(t,\tau)\Big[G(\tau)P_vG^T(\tau) + P_w\Big]\Phi^T(t,\tau)d\tau \tag{37.21}$$

Now that we have an expression for the covariance $P(t)$ as a function of time, we need to compute its derivative $\dot{P}(t)$, in order to evaluate the Kalman filter cost function (37.8). Specifically, letting

$$\tilde{G}(t) \triangleq G(t)P_vG^T(t) + P_w \tag{37.22}$$

and using Leibniz's integration rule (Fact 2.5), we find that

$$\begin{aligned}\dot{P}(t) = {} & \dot{\Phi}(t,t_0)P(t_0)\Phi^T(t,t_0) + \Phi(t,t_0)P(t_0)\dot{\Phi}^T(t,t_0)\\ & + \int_{t_0}^{t}\Big[\dot{\Phi}(t,\tau)\tilde{G}(\tau)\Phi^T(t,\tau) + \Phi(t,\tau)\tilde{G}(\tau)\dot{\Phi}^T(t,\tau)\Big]d\tau\\ & + \Phi(t,t)\tilde{G}(t)\Phi^T(t,t)\end{aligned} \tag{37.23}$$

where $\Phi^T(t,t) = I$. Hence combining (37.11), and (37.21)-(37.23), we find that

$$\begin{aligned}\dot{P}(t) = {} & \Big[A(t) - G(t)C(t)\Big]P(t) + P(t)\Big[A(t) - G(t)C(t)\Big]^T\\ & + G(t)P_vG^T(t) + P_w\end{aligned} \tag{37.24}$$

Finally, differentiating the cost function (37.8) with respect to $G(t)$ (see Appendix E), we find that

$$\frac{\partial J(t)}{\partial G(t)} = -2C(t)P(t) + 2P_v G^T(t) \tag{37.25}$$

Therefore, from the first-order necessary conditions of optimality, we find that the Kalman gain, that is, the gain $G(t)$ which minimizes (37.3) subject to Assumptions 37.1-37.3, is given by

$$G(t) = P(t)C^T(t)P_v^{-1} \tag{37.26}$$

Hence the Kalman filter estimate $\hat{x}(t)$ of $x(t)$ is the solution of

$$\begin{aligned} \dot{\hat{x}}(t) &= A(t)\hat{x}(t) + B(t)u(t) \\ &\quad + P(t)C^T(t)P_v^{-1}\Big[y(t) - C(t)\hat{x}(t) - D(t)u(t)\Big] \end{aligned} \tag{37.27}$$

where the estimation error's covariance $P(t)$ is the solution of

$$\dot{P}(t) = A(t)P(t) + P(t)A^T(t) - P(t)C^T(t)P_v^{-1}C(t)P(t) + P_w \tag{37.28}$$

Remark 37.4 To initialize the Kalman filter, we need to provide an initial estimate $\hat{x}(t_0)$ of the state $x(t_0)$, and an initial estimate $P(t_0)$ of the error covariance, which the Kalman filter gives no guidance on how to choose. From these initial values, the estimate $\hat{x}(t)$ is obtained by numerically integrating (37.27) and (37.28), where we need to integrate (37.28), since $\hat{x}(t)$ is a function of $P(t)$. ◻

37.1 Algorithm Summary

`Underlying System`

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

`Known/Measured Signals`

$$u(t),\ y(t)$$

`Assumptions`

There exists a positive semi-definite $P_w \in \mathbb{R}^{n\times n}$ and $P_v \in \mathbb{R}^{p\times p}$, such that for all $t, \tau \in \mathbb{R}$, $A(t)$, $C(t)$, and $G(t)$ are deterministic, and

$$\mathbb{E}\Big[\ w(t)e^T(t_0)\ \Big] = 0$$
$$\mathbb{E}\Big[\ v(t)e^T(t_0)\ \Big] = 0$$
$$\mathbb{E}\Big[\ w(t)w^T(\tau)\ \Big] = \delta(t-\tau)P_w$$
$$\mathbb{E}\Big[\ v(t)v^T(\tau)\ \Big] = \delta(t-\tau)P_v$$
$$\mathbb{E}\Big[\ v(t)w^T(\tau)\ \Big] = 0$$

where $e(t) \triangleq \hat{x}(t) - x(t)$, and δ denotes the Dirac delta function.

`Known Values`

$$A(t),\ B(t),\ C(t),\ D(t),\ P_w,\ P_v$$

`Estimated/Computed Signals`

$$\hat{x}(t),\ P(t)$$

`User-Supplied Initial Conditions`

$$\hat{x}(t_0),\ P(t_0)$$

`Estimate and Covariance Dynamics`

$$\begin{aligned}\dot{\hat{x}}(t) &= A(t)\hat{x}(t) + B(t)u(t)\\ &\quad + P(t)C^T(t)P_v^{-1}\Big[y(t) - C(t)\hat{x}(t) - D(t)u(t)\Big]\\ \dot{P}(t) &= A(t)P(t) + P(t)A^T(t) - P(t)C^T(t)P_v^{-1}C(t)P(t) + P_w\end{aligned}$$

37.2 Comments and Observations

There are some things that were ignored in the derivation of the Kalman filter which the reader should be aware of:

(1) The process and measurement noise do not need to be white. If they are the outputs of known linear systems driven by white noise, then the noise dynamics can be incorporated into $A(t)$. See Section 39.

(2) The covariance matrices P_w and P_v must be known exactly. Otherwise, the Kalman filter will perform sub-optimally.

(3) The Kalman filter minimizes the trace of $\dot{P}$, which from Appendix E, is equivalent to the sum of the eigenvalues of $\dot{P}$. This does not mean that all of eigenvalues of $\dot{P}$ are simultaneously minimized. For instance, it could happen that one or more of the eigenvalues are actually increasing while the others are decreasing.

(4) $P(t)$ should be obtained by integrating (37.28) from the initial covariance matrix $P(t_0)$. However, when the system (37.1) is time-invariant, it is common to use the steady-state value P_{ss} of P to compute the gain G in (37.26), where P_{ss} is the solution of the *continuous-time algebraic Riccati (CARE) equation*

$$0 = AP_{ss} + P_{ss}A^T - P_{ss}C^T P_v^{-1} C P_{ss} + P_w \tag{37.29}$$

In this case, the gain G is a constant and must only be calculated once, specifically,

$$G = P_{ss}C^T P_v^{-1} \tag{37.30}$$

However, this filter will perform suboptimally.

(5) If $P_w \neq 0$, then the steady-state value of $P(t)$ is not zero. Furthermore, if P is initialized to a value less than the steady state value, it will actually increase until that value is achieved.

(6) The process and measurement noise do not need to be uncorrelated, to derive the Kalman filter. However, in this case, the covariance matrix $P(t)$ has some additional terms.

(7) If A, C, P_v, and P_w are known *a priori* as functions of time, then $P(t)$ does not need to be computed in real-time; it can be precomputed.

SECTION 38

Discrete-Time Kalman filter

Consider the time-varying, discrete-time linear system

$$\begin{aligned} x(k+1) &= A(k)x(k) + B(k)u(k) + w(k) \\ y(k) &= C(k)x(k) + D(k)u(k) + v(k) \end{aligned} \tag{38.1}$$

where

- $x \in \mathbb{R}^n$ denotes an unknown state vector.
- $u \in \mathbb{R}^m$ denotes a known (or measured) input.
- $y \in \mathbb{R}^p$ denotes a known (or measured) output.
- $w \in \mathbb{R}^n$ denotes an unknown process noise vector.
- $v \in \mathbb{R}^p$ denotes an unknown measurement noise vector.
- $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$, and $D \in \mathbb{R}^{p\times m}$ are known.

Then the purpose of a discrete-time Kalman filter is to produce an estimate $\hat{x}(k)$ of $x(k)$ given the signals u and y, the system matrices (A, B, C, D), some knowledge about the noise statistics, and an initial state estimate $\hat{x}(k_0)$. Specifically, the Kalman filter estimate $\hat{x}(k)$ of $x(k)$ is given by

$$\begin{aligned} \hat{x}(k|k-1) &= A(k-1)\hat{x}(k-1) + B(k-1)u(k-1) \\ \hat{x}(k) &= \hat{x}(k|k-1) + G(k)\Big[y(k) - C(k)\hat{x}(k|k-1) - D(k)u(k)\Big] \end{aligned} \tag{38.2}$$

where $\hat{x}(k|k-1)$ denotes an intermediate prediction of $x(k)$ given all of the data until time $k-1$, and the gain $G(k)$ is chosen to minimize

$$J(k) = \mathbb{E}\left[\Big\|\hat{x}(k) - x(k)\Big\|_2^2\right] \tag{38.3}$$

Remark 38.1 The gain $G(k)$ which minimizes (38.3) is called the *Kalman gain.* ❒

To compute the Kalman gain, let $e(k)$ denote the instantaneous estimation error, that is, let

$$e(k) \triangleq \hat{x}(k) - x(k) \tag{38.4}$$

Then the Kalman filter cost function is given by

$$J(k) = \mathbb{E}\Big[e^T(k)e(k)\Big] = \texttt{tr}\bigg(\mathbb{E}\Big[e(k)e^T(k)\Big]\bigg) \tag{38.5}$$

where $\texttt{tr}(\cdot)$ denotes the matrix trace (see Appendix E). Hence letting

$$P(k) \triangleq \mathbb{E}\Big[e(k)e^T(k)\Big] \tag{38.6}$$

we have that the Kalman filter cost function is given by

$$J(k) = \texttt{tr}\Big(P(k)\Big) \tag{38.7}$$

Next, from (38.7), note that we are going to need an expression for $P(k)$ in order to calculate the Kalman gain. We accomplish this using the following procedure:

(1) Compute the error dynamics $e(k)$ from (38.1) and (38.2).

(2) Use the variation of parameters formula (2.16) from Section 2 to determine an equation for $e(k)$ as a function of $e(k_0)$.

(3) Make some assumptions about the noise processes and compute $P(k)$.

We begin with the error dynamics. Specifically, letting

$$z(k-1) \triangleq G(k)v(k) - \Big[I - G(k)C(k)\Big]w(k-1) \tag{38.8}$$

then from (38.1) and (38.2), we find that

$$e(k) = \Big[I - G(k)C(k)\Big]A(k-1)e(k-1) + z(k-1) \tag{38.9}$$

Hence the current error $e(k)$ is given in terms of the initial error $e(k_0)$ by

$$e(k) = \Phi(k, k_0)e(k_0) + \sum_{\kappa=k_0}^{k-1} \Phi(k, \kappa+1)z(\kappa) \tag{38.10}$$

where the state transition matrix $\Phi(k, k_0)$ is the solution of

$$\Phi(k, k_0) = \Big[I - G(k)C(k)\Big]A(k-1)\Phi(k-1, k_0) \tag{38.11}$$

subject to the initial condition $\Phi(k_0, k_0) = I$.

Next, note that the covariance matrix $P(k)$ is given by

$$\begin{aligned} P(k) = \mathbb{E}\Bigg[&\left(\Phi(k, k_0)e(k_0) + \sum_{\kappa=k_0}^{k-1} \Phi(k, \kappa+1)z(\kappa)\right) \\ &\left(\Phi(k, k_0)e(k_0) + \sum_{\eta=k_0}^{k-1} \Phi(k, \eta+1)z(\eta)\right)^T\Bigg] \end{aligned} \tag{38.12}$$

where the state-transition matrices Φ must remain inside the expectation operator since they might be correlated with the noise processes and initial estimation error $e(k_0)$. Hence to simplify the covariance equation, we introduce the assumption:

Assumption 38.1 For all $k \in \mathbb{Z}$, $A(k)$, $C(k)$, and $G(k)$ are deterministic. □

From Assumption 38.1, it follows that the state-transition matrix $\Phi(k, k_0)$ is deterministic since (38.11) only involves deterministic matrices. Therefore, using Assumption 38.1, we find that

$$P(k) = \Phi(k, k_0)P(k_0)\Phi^T(k, k_0) + \tilde{P}(k) + \tilde{P}^T(k) + \tilde{\tilde{P}}(k) \tag{38.13}$$

where $\tilde{\tilde{P}}(k)$ contains the product of the two summation terms, and

$$\tilde{P}(k) \triangleq \sum_{\kappa=k_0}^{k-1} \Phi(k, \kappa+1)\mathbb{E}\Big[z(\kappa)e^T(k_0)\Big]\Phi^T(k, k_0) \tag{38.14}$$

Remark 38.2 Technically, $P(k)$ should only be called the covariance matrix if $\mathbb{E}[e(k)] = 0$ for all $k \in \mathbb{Z}$. From (38.10) and Assumption 38.1, we see that this is indeed the case if, for all $k \in \mathbb{Z}$:

$$\mathbb{E}\big[e(k_0)\big] = 0, \quad \mathbb{E}\big[v(k)\big] = 0, \quad \mathbb{E}\big[w(k)\big] = 0 \tag{38.15}$$

However, since these assumptions are not required for the rest of the development, we do not explicitly require them. Just be aware that without (38.15), $P(k)$ is not technically the *covariance* matrix. □

At this point, we clearly need to make some assumptions about the terms $v(\kappa+1)e^T(k_0)$ and $w(\kappa)e^T(k_0)$ to be able to evaluate $\tilde{P}(k)$. The standard assumption is that the noise processes are uncorrelated with the initial estimation error, that is:

Assumption 38.2 For all $k \in \mathbb{Z}$,

$$\mathbb{E}\Big[w(k)e^T(k_0)\Big] = 0, \qquad \mathbb{E}\Big[v(k)e^T(k_0)\Big] = 0 \tag{38.16}$$

□

Specifically, using Assumptions 38.1-38.2, we find that $\tilde{P}(k) = 0$ and

$$P(k) = \Phi(k, k_0)P(k_0)\Phi^T(k, k_0) + \tilde{\tilde{P}}(k) \tag{38.17}$$

where $\tilde{\tilde{P}}(k)$ is given by

$$\begin{aligned} \tilde{\tilde{P}}(k) &= \sum_{\kappa=k_0}^{k-1} \Phi(k, \kappa+1) \sum_{\eta=k_0}^{k-1} \mathbb{E}\Big[z(\kappa)z^T(\eta)\Big]\Phi^T(k, \eta+1) \\ &= \sum_{\kappa=k_0+1}^{k} \Phi(k, \kappa) \sum_{\eta=k_0+1}^{k} \mathbb{E}\Big[z(\kappa-1)z^T(\eta-1)\Big]\Phi^T(k, \eta) \end{aligned} \tag{38.18}$$

The final assumptions we need to make concern the terms of the form $z(\kappa-1)z^T(\eta-1)$. Specifically, since $z(\kappa-1)$ is given by

$$z(\kappa-1) \triangleq G(\kappa)v(\kappa) - \Big[I - G(\kappa)C(\kappa)\Big]w(\kappa-1) \tag{38.19}$$

where $G(\kappa)$ and $C(\kappa)$ are deterministic, this amounts to making assumptions about the products $v(\kappa)v^T(\eta)$, $v(\kappa)w^T(\eta-1)$, and $w(\kappa-1)w^T(\eta-1)$. In the standard Kalman filter, one assumes that w and v are stationary white noise processes that are uncorrelated with each other, that is:

Assumption 38.3 There exists a positive semi-definite $P_w \in \mathbb{R}^{n\times n}$ and $P_v \in \mathbb{R}^{p\times p}$, such that for all $k, \kappa \in \mathbb{Z}$:

$$\begin{aligned} \mathbb{E}\Big[\ w(k)w^T(\kappa)\ \Big] &= \delta(k-\kappa)P_w \\ \mathbb{E}\Big[\ v(k)v^T(\kappa)\ \Big] &= \delta(k-\kappa)P_v \\ \mathbb{E}\Big[\ v(k)w^T(\kappa)\ \Big] &= 0 \end{aligned} \tag{38.20}$$

where δ denotes the Dirac delta function. □

Hence letting

$$H(k) \triangleq I - G(k)C(k) \tag{38.21}$$

and using Assumptions 38.1-38.3, we find that

$$\mathbb{E}\Big[z(\kappa-1)z^T(\eta-1)\Big] = \delta(\kappa-\eta)\Big[G(\kappa)P_vG^T(\eta) + H(\kappa)P_wH^T(\eta)\Big]$$

and therefore

$$\begin{aligned} P(k) &= \Phi(k,k_0)P(k_0)\Phi^T(k,k_0) \\ &\quad + \sum_{\kappa=k_0+1}^{k} \Phi(k,\kappa)\Big[G(\kappa)P_vG^T(\kappa) + H(\kappa)P_wH^T(\kappa)\Big]\Phi^T(k,\kappa) \end{aligned} \tag{38.22}$$

Next, we develop an expression for the covariance $P(k)$ in terms of $P(k-1)$ by using the state-transition matrix (38.11). Specifically, substituting (38.11) into (38.22), and letting

$$\tilde{A}(k-1) \triangleq A(k-1)P(k-1)A^T(k-1) + P_w \tag{38.23}$$

we find that

$$P(k) = \Big[I - G(k)C(k)\Big]\tilde{A}(k-1)\Big[I - G(k)C(k)\Big]^T + G(k)P_vG^T(k) \tag{38.24}$$

Finally, differentiating the cost function (38.7) with respect to $G(k)$ (see Appendix E), we find that

$$\frac{\partial J(k)}{\partial G(k)} = -2C(k)\tilde{A}(k-1) + 2\Big[C(k)\tilde{A}(k-1)C^T(k) + P_v\Big]G^T(k) \tag{38.25}$$

Therefore, from the first-order necessary conditions of optimality, we find that the Kalman gain, that is, the gain $G(k)$ which minimizes (38.3) subject to Assumptions 38.1-38.3, is given by

$$G(k) = \tilde{A}(k-1)C^T(k)\Big[C(k)\tilde{A}(k-1)C^T(k) + P_v\Big]^{-1} \tag{38.26}$$

Hence the Kalman filter estimate $\hat{x}(k)$ of $x(k)$ is given by (38.2), where $G(k)$ is given by (38.26), $\tilde{A}(k-1)$ is given by (38.23), and

$$P(k) = \Big[I - G(k)C(k)\Big]\Big[A(k-1)P(k-1)A^T(k-1) + P_w\Big] \tag{38.27}$$

38.1 Algorithm Summary

`Underlying System`

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

`Known/Measured Signals`

$$u(k),\ y(k)$$

`Assumptions`

There exists a positive semi-definite $P_w \in \mathbb{R}^{n\times n}$ and $P_v \in \mathbb{R}^{p\times p}$, such that for all $k, \kappa \in \mathbb{Z}$, $A(k)$, $C(k)$, and $G(k)$ are deterministic, and

$$\begin{aligned} &\mathbb{E}\Big[\ w(k)e^T(k_0)\ \Big] = 0, && \mathbb{E}\Big[\ w(k)w^T(\kappa)\ \Big] = \delta(k-\kappa)P_w \\ &\mathbb{E}\Big[\ v(k)e^T(k_0)\ \Big] = 0, && \mathbb{E}\Big[\ v(k)v^T(\kappa)\ \Big] = \delta(k-\kappa)P_v \\ &\mathbb{E}\Big[\ v(k)w^T(\kappa)\ \Big] = 0 \end{aligned}$$

where $e(k) \triangleq \hat{x}(k) - x(k)$, and δ denotes the Dirac delta function.

`Known Values`

$$A(k),\ B(k),\ C(k),\ D(k),\ P_w,\ P_v$$

`Estimated/Computed Signals`

$$\hat{x}(k),\ P(k)$$

`User-Supplied Initial Conditions`

$$\hat{x}(k_0),\ P(k_0)$$

`Estimate and Covariance Dynamics`

$$\begin{aligned} \tilde{A}(k-1) &= A(k-1)P(k-1)A^T(k-1) + P_w \\ G(k) &= \tilde{A}(k-1)C^T(k)\left[C(k)\tilde{A}(k-1)C^T(k) + P_v\right]^{-1} \\ P(k) &= \Big[I - G(k)C(k)\Big]\tilde{A}(k-1) \\ \hat{x}(k|k-1) &= A(k-1)\hat{x}(k-1) + B(k-1)u(k-1) \\ \hat{x}(k) &= \hat{x}(k|k-1) + G(k)\Big[y(k) - C(k)\hat{x}(k|k-1) - D(k)u(k)\Big] \end{aligned}$$

SECTION 39

Handling Colored Noise

Consider the time-varying, discrete-time state-space system

$$\begin{aligned} x(k+1) &= A(k)x(k) + B(k)u(k) + w(k) \\ y(k) &= C(k)x(k) + D(k)u(k) + v(k) \end{aligned} \tag{39.1}$$

where the process noise $w \in \mathbb{R}^m$ and measurement noise $v \in \mathbb{R}^p$ are colored noise processes. Specifically, let w and v be given by

$$\begin{aligned} x_w(k+1) &= A_w(k)x_w(k) + \underline{w}(k) \\ w(k) &= C_w(k)x_w(k) \\ x_v(k+1) &= A_v(k)x_v(k) + \underline{v}(k) \\ v(k) &= C_v(k)x_v(k) \end{aligned} \tag{39.2}$$

where $\underline{w}$ and $\underline{v}$ are white noise processes, that is, there exist $P_{\underline{w}}$ and $P_{\underline{v}}$ such that for all $k, \kappa \in \mathbb{Z}$:

$$\begin{aligned} \mathbb{E}\Big[\ \underline{w}(k)\underline{w}^T(\kappa)\ \Big] &= \delta(k-\kappa)P_{\underline{w}} \\ \mathbb{E}\Big[\ \underline{v}(k)\underline{v}^T(\kappa)\ \Big] &= \delta(k-\kappa)P_{\underline{v}} \\ \mathbb{E}\Big[\ \underline{v}(k)\underline{w}^T(\kappa)\ \Big] &= 0 \end{aligned} \tag{39.3}$$

Then letting

$$\begin{aligned} \tilde{A}(k) &\triangleq \begin{bmatrix} A(k) & C_w(k) & 0 \\ 0 & A_w(k) & 0 \\ 0 & 0 & A_v(k) \end{bmatrix} \\ \tilde{C}(k) &\triangleq \begin{bmatrix} C(k) & 0 & C_v(k) \end{bmatrix} \\ \tilde{B}(k) &\triangleq \begin{bmatrix} B(k) \\ 0 \\ 0 \end{bmatrix}, \quad \tilde{x}(k) \triangleq \begin{bmatrix} x(k) \\ x_w(k) \\ x_v(k) \end{bmatrix}, \quad \tilde{w}(k) \triangleq \begin{bmatrix} 0 \\ \underline{w}(k) \\ \underline{v}(k) \end{bmatrix} \end{aligned} \tag{39.4}$$

we find that

$$\begin{aligned} \tilde{x}(k+1) &= \tilde{A}(k)\tilde{x}(k) + \tilde{B}(k)u(k) + \tilde{w}(k) \\ y(k) &= \tilde{C}(k)\tilde{x}(k) + D(k)u(k) \end{aligned} \tag{39.5}$$

where $\tilde{w}$ is a white-noise process, that is, for all $k, \kappa \in \mathbb{Z}$:

$$\mathbb{E}\Big[\tilde{w}(k)\tilde{w}^T(\kappa)\Big] = \delta(k-\kappa)\begin{bmatrix} 0 & 0 & 0 \\ 0 & P_{\underline{w}} & 0 \\ 0 & 0 & P_{\underline{v}} \end{bmatrix} \tag{39.6}$$

Hence the augmented state-space model (39.5) behaves just like the standard time-varying, discrete-time state-space model (38.1) with white process noise, and no measurement noise, that is, $v(k) = 0$. Therefore the noise in the augmented model (39.5) satisfies Assumption 38.3, where P_w is given by (39.6), and $P_v = 0$. If, in addition, we find that Assumptions 38.1 and 38.2 hold for the augmented model (39.5), then we can use the discrete-time Kalman filter presented in Section 38 to estimate the state $\tilde{x}$ of the augmented model (39.5).

Remark 39.1 The state x of the true system (39.1), along with the internal states x_w and x_v of the noise dynamics, are sub-vectors of the augmented state $\tilde{x}$. Hence an estimate $\hat{x}$ of x can be easily extracted from the Kalman filter estimate $\hat{\tilde{x}}$ of $\tilde{x}$. □

Remark 39.2 The process for handling colored noise that we presented in this section can also be applied to continuous-time systems with the appropriate modifications. However, there is one glaring problem: In the continuous-time case, the Kalman gain $G(t)$ is a function of P_v^{-1} (see equation (37.26)), although in the colored-measurement noise case, $P_v = 0$. Hence in the colored-measurement noise case, P_v^{-1} is not defined.

In practice, this problem is usually circumvented by choosing a sufficiently small value of P_v. However, this should not bother you too much since you will rarely see a continuous-time Kalman filter implemented anyway. □

SECTION 40

Discrete-Time Extended Kalman Filter

Consider the time-varying, discrete-time nonlinear system

$$\begin{aligned} x(k+1) &= f\Big(x(k), u(k)\Big) + w(k) \\ y(k) &= h\Big(x(k), u(k)\Big) + v(k) \end{aligned} \tag{40.1}$$

where

- $x \in \mathbb{R}^n$ denotes an unknown state vector.
- $u \in \mathbb{R}^m$ denotes a known (or measured) input.
- $y \in \mathbb{R}^p$ denotes a known (or measured) output.
- $w \in \mathbb{R}^n$ denotes an unknown process noise vector.
- $v \in \mathbb{R}^p$ denotes an unknown measurement noise vector.
- $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ and $h : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^p$ are known functions.

Then the purpose of a discrete-time extended Kalman filter is to produce an estimate $\hat{x}(k)$ of $x(k)$, given the signals u and y, the functions f and h, some knowledge about the noise statistics, and an initial state estimate $\hat{x}(k_0)$. Specifically, the extended Kalman filter estimate $\hat{x}(k)$ of $x(k)$ is given by

$$\begin{aligned} \hat{x}(k|k-1) &= f\Big(\hat{x}(k-1), u(k-1)\Big) \\ \hat{x}(k) &= \hat{x}(k|k-1) + G(k)\Big[y(k) - h\Big(\hat{x}(k|k-1), u(k)\Big)\Big] \end{aligned} \tag{40.2}$$

where $\hat{x}(k|k-1)$ can be thought of as an intermediate prediction of $x(k)$ given all of the data until time $k-1$.

Next, as in the linear case, we would like to choose the gain $G(k)$ which minimizes

$$J(k) = \mathbb{E}\Big[e(k)^T e(k)\Big] \tag{40.3}$$

where $e(k)$ denotes the estimation error, that is,

$$e(k) \triangleq \hat{x}(k) - x(k) \tag{40.4}$$

However, due to the nonlinearities f and h, computing the value of $G(k)$ which minimizes (40.3) is, in general, quite difficult. Instead we will compute a linearization of the estimation error $e(k)$, and show that its dynamics are similar to those of a time-varying linear system, thereby allowing us to use the results of the previous section.

To accomplish this, first note that from (40.2), the estimation error is equivalently given by

$$\begin{aligned} e(k) = f\Big(\hat{x}(k-1), u(k-1)\Big) - f\Big(x(k-1), u(k-1)\Big) - w(k-1) \\ + G(k)\Big[h\Big(x(k), u(k)\Big) + v(k) - h\Big(\hat{x}(k|k-1), u(k)\Big)\Big] \end{aligned} \tag{40.5}$$

Next, note that if $\hat{x}(k-1)$ is "close" to $x(k-1)$, then $f\big(x(k-1), u(k-1)\big)$ can be approximated well by its first-order estimate, that is, linearizing $f\big(x(k-1), u(k-1)\big)$ about $\hat{x}(k-1)$, we find that

$$f\Big(x(k-1), u(k-1)\Big) \approx f\Big(\hat{x}(k-1), u(k-1)\Big) - A(k-1)e(k-1) \tag{40.6}$$

where

$$A(k-1) \triangleq \left.\frac{\partial f(x,u)}{\partial x}\right|_{\hat{x}(k-1), u(k-1)} \tag{40.7}$$

Furthermore, if the intermediate estimate $\hat{x}(k|k-1)$ is "close" to $x(k)$, then $h\big(x(k), u(k)\big)$ can also be approximated well by its first-order estimate, that is, linearizing $h\big(x(k), u(k)\big)$ about $\hat{x}(k|k-1)$, we find that

$$h\Big(x(k), u(k)\Big) \approx h\Big(\hat{x}(k|k-1), u(k)\Big) + C(k)\Big[x(k) - \hat{x}(k|k-1)\Big] \tag{40.8}$$

where

$$C(k) \triangleq \left.\frac{\partial h(x,u)}{\partial x}\right|_{\hat{x}(k|k-1), u(k)} \tag{40.9}$$

Specifically, since

$$\begin{aligned} x(k) - \hat{x}(k|k-1) = f\Big(x(k-1), u(k-1)\Big) + w(k-1) \\ - f\Big(\hat{x}(k-1), u(k-1)\Big) \end{aligned} \tag{40.10}$$

then using the first order approximation (40.6) of $f\big(x(k-1), u(k-1)\big)$ in (40.8), it follows that

$$h\Big(x(k), u(k)\Big) \approx h\Big(\hat{x}(k|k-1), u(k)\Big) - C(k)\Big[A(k-1)e(k-1) - w(k-1)\Big] \tag{40.11}$$

Hence using the approximations (40.6) and (40.11) in (40.5), we find that the estimation error is approximately given by

$$e(k) \approx \Big[I - G(k)C(k)\Big]\Big[A(k-1)e(k-1) - w(k-1)\Big] + G(k)v(k) \tag{40.12}$$

Surprisingly, the equation (40.12) describing the approximate estimation error dynamics is exactly the same as in the linear case (38.9). Thus if Assumptions (38.1)-(38.3) hold, then the gain $G(k)$ which minimizes (40.3) is approximately given by (38.26) and the covariance matrix $P(k)$ of the estimation error is approximately given by (38.27), where $A(k)$ and $C(k)$ are given by (40.7) and (40.9), respectively.

There are two additional issues that are unique to the nonlinear case:

1) Since $P(k)$ is a covariance matrix, it must be a positive semi-definite matrix. However, since we are approximating the dynamics with a linearization, it could happen that using equation (38.27), $P(k)$ becomes an indefinite matrix. In this case, the user must reset $P(k)$ to a positive semi-definite matrix. This process is sometimes called *covariance resetting.*

2) As in the linear case, we use Assumptions (38.1)-(38.3). However, the user should be aware that since the minimization of (40.3) is only approximate in the nonlinear case, there could exist instances where removing one of the assumptions actually improves the filter's performance.

40.1 Algorithm Summary

Underlying System

$$x(k+1) = f\Big(x(k), u(k)\Big) + w(k)$$

$$y(k) = h\Big(x(k), u(k)\Big) + v(k)$$

Assumptions

There exists a positive semi-definite $P_w \in \mathbb{R}^{n \times n}$ and $P_v \in \mathbb{R}^{p \times p}$, such that for all $k, \kappa \in \mathbb{Z}$, f, h, and $G(k)$ are deterministic, and

$$\mathbb{E}\Big[\ w(k)e^T(k_0)\ \Big] = 0, \qquad \mathbb{E}\Big[\ w(k)w^T(\kappa)\ \Big] = \delta(k-\kappa)P_w$$

$$\mathbb{E}\Big[\ v(k)e^T(k_0)\ \Big] = 0, \qquad \mathbb{E}\Big[\ v(k)v^T(\kappa)\ \Big] = \delta(k-\kappa)P_v$$

$$\mathbb{E}\Big[\ v(k)w^T(\kappa)\ \Big] = 0$$

where $e(k) \triangleq \hat{x}(k) - x(k)$, and δ denotes the Dirac delta function.

Known or Measured Signals, Functions, and Values

$u(k)$, $y(k)$, f, h, P_w, P_v

Estimated/Computed Signals

$\hat{x}(k)$, $P(k)$

User-Supplied Initial Conditions

$\hat{x}(k_0)$, $P(k_0)$

Estimate and Covariance Dynamics

$$A(k-1) \triangleq \left.\frac{\partial f(x,u)}{\partial x}\right|_{\hat{x}(k-1),u(k-1)}$$

$$\tilde{A}(k-1) = A(k-1)P(k-1)A^T(k-1) + P_w$$

$$\hat{x}(k|k-1) = f\Big(\hat{x}(k-1), u(k-1)\Big)$$

$$C(k) \triangleq \left.\frac{\partial h(x,u)}{\partial x}\right|_{\hat{x}(k|k-1),u(k)}$$

$$G(k) = \tilde{A}(k-1)C^T(k)\Big[C(k)\tilde{A}(k-1)C^T(k) + P_v\Big]^{-1}$$

$$P(k) = \Big[I - G(k)C(k)\Big]\tilde{A}(k-1)$$

$$\hat{x}(k) = \hat{x}(k|k-1) + G(k)\Big[y(k) - h\Big(\hat{x}(k|k-1), u(k)\Big)\Big]$$

SECTION 41

Extended Kalman Filter-Based Identification

Having spent a considerable amount of time on Kalman filters, it is natural to wonder if we can use them for system identification. On the surface, it may appear that state estimation and system identification have nothing in common. However, as we show here, we can indeed contort the extended Kalman filter to be used as a system identification scheme. Specifically, consider the time-varying, discrete-time linear system

$$\begin{aligned} x(k+1) &= A(k)x(k) + B(k)u(k) + w(k) \\ y(k) &= C(k)x(k) + D(k)u(k) + v(k) \end{aligned} \tag{41.1}$$

where

$$\begin{aligned} A(k+1) &= f_A\Big(A(k)\Big), & B(k+1) &= f_B\Big(B(k)\Big) \\ C(k+1) &= f_C\Big(C(k)\Big), & D(k+1) &= f_D\Big(D(k)\Big) \end{aligned} \tag{41.2}$$

and

- $x \in \mathbb{R}^n$ denotes an unknown state vector.
- $u \in \mathbb{R}^m$ denotes a known (or measured) input.
- $y \in \mathbb{R}^p$ denotes a known (or measured) output.
- $w \in \mathbb{R}^n$ denotes an unknown process noise vector.
- $v \in \mathbb{R}^p$ denotes an unknown measurement noise vector.
- $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$, and $D \in \mathbb{R}^{p\times m}$ are unknown.
- f_A, f_B, f_C, and f_D are known.

Remark 41.1 The difference between this scenario and the standard discrete-time Kalman filter scenario (see Section 38) is that the system matrices (A, B, C, D) are now considered to be unknown. □

Remark 41.2 If the system 43.1 is time-invariant, then for all $k \in \mathbb{Z}$,

$$\begin{aligned} A(k+1) &= A(k), & B(k+1) &= B(k) \\ C(k+1) &= C(k), & D(k+1) &= D(k) \end{aligned} \tag{41.3}$$

□

Next, consider the augmented matrices

$$f\Big(\tilde{x}(k), u(k)\Big) \triangleq \begin{bmatrix} A(k)x(k) \\ \texttt{vec}\Big(f_A\big(A(k)\big)\Big) \\ \texttt{vec}\Big(f_B\big(B(k)\big)\Big) \\ \texttt{vec}\Big(f_C\big(C(k)\big)\Big) \\ \texttt{vec}\Big(f_D\big(D(k)\big)\Big) \end{bmatrix}, \quad \tilde{x}(k) \triangleq \begin{bmatrix} x(k) \\ \texttt{vec}\Big(A(k)\Big) \\ \texttt{vec}\Big(B(k)\Big) \\ \texttt{vec}\Big(C(k)\Big) \\ \texttt{vec}\Big(D(k)\Big) \end{bmatrix} \tag{41.4}$$

$$h\Big(\tilde{x}(k), u(k)\Big) \triangleq C(k)x(k) + D(k)u(k)$$

where `vec` denotes the vectorization operator (see Appendix F). Then

$$\begin{aligned} \tilde{x}(k+1) &= f\Big(\tilde{x}(k), u(k)\Big) + w(k) \\ y(k) &= h\Big(\tilde{x}(k), u(k)\Big) + v(k) \end{aligned} \tag{41.5}$$

and hence the augmented system (43.5) is of the form required by the discrete-time extended Kalman filter which we developed in Section 40. Specifically, using the discrete-time extended Kalman filter, we can estimate the augmented state $\tilde{x}(k)$, which contains the system matrices (A, B, C, D). Therefore we can think of the discrete-time extended Kalman filter applied to our augemented system (43.5) as a system identification technique for our original state-space model (43.1).

PART XII

Subspace Identification

In Section 41, we showed how a discrete-time extended Kalman filter can be used for system identification. However, since the extended Kalman filter only yields an approximate solution to the nonlinear Kalman cost, it is worth considering some of the problems with this approach. Specifically, consider again the the time-varying discrete-time model

$$\begin{aligned} x(k+1) &= A(k)x(k) + B(k)u(k) + w(k) \\ y(k) &= C(k)x(k) + D(k)u(k) + v(k) \end{aligned} \tag{41.6}$$

which we introduced in Section 41.

The identification of (41.6) is fundamentally difficult since the model (41.6) is not linear in the unknowns x, A, B, C, and D. Specifically, due to the products $A(k)x(k)$ and $C(k)x(k)$, we say that the model (41.6) is *bilinear* in the unknowns. The following example demonstrates one difficultly that may arise in estimating a bilinear system, or a more general nonlinear system:

Example 41.1 Consider the system

$$y(k) = \alpha\beta \tag{41.7}$$

where $\alpha, \beta \in \mathbb{R}$ are unknown, but y is measured for $k = 1, \ldots, N$.
Q: Is it possible to identify α and β exactly given the measurements of y?
A: No. We can only identify the product $\alpha\beta$.

For instance, suppose that $\alpha = 1$ and $\beta = 2$. Then for all $k = 1, \ldots, N$, $y(k) = 2$. Hence, an equally valid model of the system is given by

$$y(k) = \hat{\alpha}\hat{\beta} \tag{41.8}$$

where $\hat{\alpha} = \frac{1}{4}$ and $\hat{\beta} = 8$. In fact, there are an infinite number of solutions $(\hat{\alpha}, \hat{\beta})$ capable of exactly producing the output y.

Example 41.1 demonstrates that a bilinear model may have an infinite number of equally valid representations, regardless of how the input u is chosen, that is, the identifiability of the model has nothing to do with the persistency of the input. For instance, there does not exist an input u for which the coefficients α and β in the system

$$y(k) = \alpha\beta u(k) \tag{41.9}$$

can be uniquely identified. Therefore, we say that the system (41.9) is *poorly defined* or *not identifiable.*

Example 41.1 also raises a more general point about our extended Kalman filter identification scheme. Specifically, since the state-space model (41.6) is bilinear in the unknowns, how do we know if there exists a unique solution to the Kalman cost function (40.3) for the augmented model (43.5)?

To answer this question, consider the model (41.6) with no noise, that is, consider the state-space model

$$\begin{aligned} x(k+1) &= A(k)x(k) + B(k)u(k) \\ y(k) &= C(k)x(k) + D(k)u(k) \end{aligned} \tag{41.10}$$

Then from Section 4, it follows that applying a similarity transform to (41.10) yields an equivalent state-space model, that is, for all invertible matrices $P \in \mathbb{R}^{n \times n}$, the state-space model (41.10) is equivalent to the model

$$\begin{aligned} x_p(k+1) &= A_p(k)x_p(k) + B_p(k)u(k) \\ y(k) &= C_p(k)x_p(k) + D_p(k)u(k) \end{aligned} \tag{41.11}$$

where

$$\begin{aligned} x_p(k) = Px(k), \quad A_p(k) &= PA(k)P^{-1}, \quad & B_p(k) &= PB(k) \\ C_p(k) &= C(k)P^{-1}, & D_p(k) &= D(k) \end{aligned} \tag{41.12}$$

Hence if neither the state x, nor the system matrices (A, B, C, D) are fixed, then there exist an infinite number of equally valid state-space representations of (41.10). In this sense, state-space models behave a lot like the model (41.7) in Example 41.1, that is, when neither the state x, nor the system matrices (A, B, C, D) are fixed, then we have little chance of being able to identify the model uniquely.

In spite of this hurdle, state-space system identification is alive and well in the form of *subspace identification* algorithms. The trick, as it turns out, is to identify the system matrices (A, B, C, D) indirectly. Specifically, instead of attempting to identify the system matrices (A, B, C, D) directly, subspace identification algorithms attempt to identify a subspace of the input/output data which can easily be converted back into the system matrices (A, B, C, D). This discussion can get very esoteric very quickly, so let's jump right into one of the first, and most well-known subspace identification algorithms: OKID.

SECTION 42

Observer/Kalman Filter Identification (OKID)

Consider the time-invariant, discrete-time linear system

$$\begin{aligned} x(k+1) &= Ax(k) + Bu(k) + w(k) \\ y(k) &= Cx(k) + Du(k) + v(k) \end{aligned} \tag{42.1}$$

where

- $x \in \mathbb{R}^n$ denotes an unknown state vector.
- $u \in \mathbb{R}^m$ denotes a known (or measured) input.
- $y \in \mathbb{R}^p$ denotes a known (or measured) output.
- $w \in \mathbb{R}^n$ denotes an unknown process noise vector.
- $v \in \mathbb{R}^p$ denotes an unknown measurement noise vector.
- $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$, and $D \in \mathbb{R}^{p\times m}$ are unknown.
- (A, C) is observable.

Then the observer/Kalman filter identification (OKID) algorithm provides a method for estimating the system matrices A, B, C, and D indirectly [19]. Specifically, OKID breaks up the estimation process into three steps:

(1) Estimate the Markov parameters $D, C\tilde{B}, \ldots, C\tilde{A}^{n-1}\tilde{B}$ of a modified state-space model.

(2) Estimate the Markov parameters $D, CB, \ldots, CA^{n-1}B$ of the true system (42.1) from the estimates of $D, C\tilde{B}, \ldots, C\tilde{A}^{n-1}\tilde{B}$.

(3) Estimate the system matrices A, B, C, and D from the Markov parameter estimates of $D, CB, \ldots, CA^{n-1}B$.

42.1 General Observations

From the discrete variation of parameters formula (2.27), we find that the output $y(k)$ of (42.1) is given in terms of the initial condition $x(0)$ by

$$y(k) = CA^k x(0) + v(k) + Du(k) + \sum_{\kappa=0}^{k-1} CA^{k-1-\kappa}\Big[Bu(\kappa) + w(\kappa)\Big] \tag{42.2}$$

Hence collecting the Markov parameters into matrices, that is, letting

$$\mathcal{H}_k(A,B,C,D) \triangleq \begin{bmatrix} D & CB & CAB & \cdots & CA^{k-1}B \end{bmatrix} \tag{42.3}$$

$$\tilde{\mathcal{H}}_k(A,B,C) \triangleq \begin{bmatrix} CB & CAB & \cdots & CA^{k-1}B \end{bmatrix} \tag{42.4}$$

then for $k \geq 1$, we find that (42.2) is equivalently written as

$$\begin{aligned} y(k) &= CA^k x(0) + v(k) \\ &\quad + \mathcal{H}_k(A,B,C,D) \begin{bmatrix} u(k) \\ \vdots \\ u(0) \end{bmatrix} + \tilde{\mathcal{H}}_k(A,I,C) \begin{bmatrix} w(k-1) \\ \vdots \\ w(0) \end{bmatrix} \end{aligned} \tag{42.5}$$

Furthermore, letting

$$\phi_{i,N}(y) \triangleq \begin{bmatrix} y(i) & \cdots & y(N) \end{bmatrix} \tag{42.6}$$

$$\mathcal{T}_{i,N}(u) \triangleq \begin{bmatrix} \underbrace{\begin{matrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{matrix}}_{i \text{ columns}} & \begin{matrix} u(0) & \cdots & u(N-i) \\ \ddots & \ddots & \vdots \\ \cdots & 0 & u(0) \end{matrix} \end{bmatrix} \tag{42.7}$$

and evaluating (42.5) at times $k = 0, \ldots, N$, we find that

$$\begin{aligned} \phi_{0,N}(y) &= \begin{bmatrix} CA^0 x(0) & \cdots & CA^N x(0) \end{bmatrix} + \phi_{0,N}(v) \\ &\quad + \mathcal{H}_N(A,B,C,D)\mathcal{T}_{0,N}(u) + \tilde{\mathcal{H}}_N(A,I,C)\mathcal{T}_{1,N}(w) \end{aligned} \tag{42.8}$$

Equation (42.8) encapsulates the regression equations for the state-space model (42.1). Unfortunately, for the purpose of estimating the matrices A, B, C, and D or their Markov parameters, it is not very useful. Specifically, since $\phi_{0,N}(y)$ and $\mathcal{T}_{0,N}(u)$ are the only matrices that we know or measure in (42.8), (42.8) is not immediately helpful for estimating either the matrices (A, B, C, D) or their Markov parameters.

Example 42.1 Suppose that the Markov parameter vector is estimated using the least-squares estimate

$$\hat{\mathcal{H}}_N(A,B,C,D) = \underset{\Theta}{\operatorname{argmin}} \left\| \phi_{0,N}(y) - \Theta \mathcal{T}_{0,N}(u) \right\|_F \tag{42.9}$$

Furthermore, assume that $\mathcal{T}_{0,N}(u)$ has full row rank. Then letting

$$\mathcal{T}_{0,N}^{+}(u) \triangleq \mathcal{T}_{0,N}^{T}(u)\Big(\mathcal{T}_{0,N}(u)\mathcal{T}_{0,N}^{T}(u)\Big)^{-1} \tag{42.10}$$

we find that

$$\begin{aligned}\hat{\mathcal{H}}_N(A,B,C,D) &= \phi_{0,N}(y)\mathcal{T}_{0,N}^{+}(u) &(42.11)\\ &= \mathcal{H}_N(A,B,C,D) + \tilde{\mathcal{H}}_N(A,I,C)\mathcal{T}_{1,N}(w)\mathcal{T}_{0,N}^{+}(u)\\ &\quad + \Big(\left[\begin{array}{ccc} CA^0x(0) & \cdots & CA^Nx(0)\end{array}\right] + \phi_{0,N}(v)\Big)\mathcal{T}_{0,N}^{+}(u)\end{aligned}$$

Hence even if the noise processes are zero, we still find that the estimate (42.9) is biased since

$$\begin{aligned}\mathbb{E}\Big[\hat{\mathcal{H}}_N(A,B,C,D)\Big] &- \mathcal{H}_N(A,B,C,D)\\ &= \left[\begin{array}{ccc} CA^0x(0) & \cdots & CA^Nx(0)\end{array}\right]\mathcal{T}_{0,N}^{+}(u)\end{aligned} \tag{42.12}$$

In fact, for the estimate (42.9) to be unbiased, we would also have to assume that $x(0) = 0$, which is quite a restrictive assumption[9]. □

42.2 Introducing a Kalman-Like Gain

The previous section showed that the regression equations (42.8) for the system (42.1) are a function of the initial condition $x(0)$. Furthermore, we showed that this dependence generally causes least-squares estimates of the Markov parameter vector $\mathcal{H}_N(A,B,C,D)$, such as (42.9), to be biased. The idea behind OKID is to introduce a matrix G which removes the initial condition $x(0)$ from the regression equations (42.8). Specifically, since this choice of G acts like the Kalman gain, the algorithm is referred to as the observer/Kalman filter identification algorithm.

[9] Technically, we only have to have that $x(0)$ is in the nullspace of A. However, since most discrete-time systems are obtained by sampling continuous-time systems, the matrix A usually has full rank, see Section 3. In this case, we require that $x(0) = 0$.

First, let G denote an arbitrary matrix in $\mathbb{R}^{n\times p}$. Furthermore, add and subtract the vector $Gy(k)$ from the state equation of (42.1), that is,

$$x(k+1) = Ax(k) + Bu(k) + w(k) + Gy(k) - Gy(k) \tag{42.13}$$

Then using the output equation of (42.1) in (42.13), we find that the system (42.1) is equivalently represented by

$$\begin{aligned} x(k+1) &= (A+GC)x(k) + (B+GD)u(k) - Gy(k) + w(k) + Gv(k) \\ y(k) &= Cx(k) + Du(k) + v(k) \end{aligned}$$

Hence the output $y(k)$ of (42.1) is given in terms of G and the initial condition $x(0)$ by

$$\begin{aligned} y(k) &= C(A+GC)^k x(0) + Du(k) + v(k) \\ &\quad + \sum_{\kappa=0}^{k-1} C(A+GC)^{k-1-\kappa}\Big[(B+GD)u(\kappa) - Gy(\kappa) + w(\kappa) + Gv(\kappa)\Big] \end{aligned} \tag{42.14}$$

Next, note that since (A, C) is controllable, we can choose G such that $(A+GC)^k = 0$ for all $k \geq n$. In fact, choosing G in this way, we find that for all $k \geq n$,

$$\begin{aligned} y(k) &= v(k) + Du(k) \\ &\quad + \sum_{\kappa=k-n}^{k-1} C(A+GC)^{k-1-\kappa}\Big[(B+GD)u(\kappa) - Gy(\kappa) + w(\kappa) + Gv(\kappa)\Big] \end{aligned} \tag{42.15}$$

and hence G has the effect of removing the initial condition $x(0)$ from the output equation. Furthermore, letting

$$\tilde{A} \triangleq A + GC, \quad \tilde{B} \triangleq \begin{bmatrix} B+GD, & -G \end{bmatrix}, \quad z(k) \triangleq \begin{bmatrix} u(k) \\ y(k) \end{bmatrix} \tag{42.16}$$

it follows that (42.15) can be concisely written as

$$y(k) = v(k) + Du(k) + \sum_{\kappa=k-n}^{k-1} C\tilde{A}^{k-1-\kappa}\Big[\tilde{B}z(\kappa) + w(\kappa) + Gv(\kappa)\Big] \tag{42.17}$$

Finally, letting

$$\Phi_{0,n-1,N}(z) \triangleq \begin{bmatrix} z(n-1) & \cdots & z(N) \\ \vdots & & \vdots \\ z(0) & \cdots & z(N-n) \end{bmatrix} \tag{42.18}$$

$$\tilde{\Phi}_{0,n,N}(u,z) \triangleq \left[\begin{array}{ccc} u(n) & \cdots & u(N) \\ \hline \multicolumn{3}{c}{\Phi_{0,n-1,N}(z)} \end{array}\right] \tag{42.19}$$

and evaluating (42.17) at times $k = n, \ldots, N$, we find that

$$\begin{aligned} \phi_{n,N}(y) &= \mathcal{H}_n(\tilde{A},\tilde{B},C,D)\tilde{\Phi}_{0,n,N}(u,z) + \tilde{\mathcal{H}}_n(\tilde{A},\tilde{B},I)\Phi_{0,n-1,N}(w) \\ &\quad + \mathcal{H}_n(\tilde{A},\tilde{B},G,I)\Phi_{0,n,N}(v) \end{aligned} \tag{42.20}$$

Therefore, estimates of the Markov parameters $D, C\tilde{B}, \ldots, C\tilde{A}^{n-1}\tilde{B}$ can be found using the least-squares estimate

$$\hat{\mathcal{H}}_n(\tilde{A},\tilde{B},C,D) = \underset{\Theta}{\operatorname{argmin}} \left\| \phi_{n,N}(y) - \Theta\tilde{\Phi}_{0,n,N}(u,z) \right\|_F \tag{42.21}$$

Remark 42.1 If $\tilde{\Phi}_{0,n,N}(u,z)$ has full row rank, then from (42.20) and (42.21), we find that

$$\begin{aligned} \hat{\mathcal{H}}_n(\tilde{A},\tilde{B},C,D) &= \phi_{n,N}(y)\tilde{\Phi}^{+}_{0,n,N}(u,z) \qquad\qquad (42.22) \\ &= \mathcal{H}_n(\tilde{A},\tilde{B},C,D) + \tilde{H}_n(\tilde{A},\tilde{B},I)\Phi_{0,n-1,N}(w)\tilde{\Phi}^{+}_{0,n,N}(u,z) \\ &\quad + H_n(\tilde{A},\tilde{B},G,I)\Phi_{0,n,N}(v)\tilde{\Phi}^{+}_{0,n,N}(u,z) \end{aligned}$$

Hence if there is no noise present, then the estimate (42.22) is an unbiased estimate of $\mathcal{H}_n(\tilde{A},\tilde{B},C,D)$, that is,

$$\mathbb{E}\left[\hat{\mathcal{H}}_n(\tilde{A},\tilde{B},C,D)\right] = \mathcal{H}_n(\tilde{A},\tilde{B},C,D) \tag{42.23}$$

This is in contrast with the least-squares estimate (42.9) of $\mathcal{H}_N(A,B,C,D)$, which has the bias (42.12). ◻

42.3 Recovering the Markov Parameters and System Matrices

The OKID algorithm introduces a matrix G which removes the initial condition $x(0)$ from the regression equations (42.15). However, the penalty for introducing G is that the OKID estimate of the Markov parameters (42.21) is now a function of the gain G. Specifically, (42.21) provides estimates of the modified Markov parameters

$$\tilde{H}_0 \triangleq D \quad \text{and} \quad \tilde{H}_i \triangleq C\tilde{A}^{i-1}\tilde{B} \quad \text{for } i \geq 1 \tag{42.24}$$

where $\tilde{A}$ and $\tilde{B}$ are given by (42.16), although the Markov parameters of the true system (42.1) are given by

$$H_0 \triangleq D \quad \text{and} \quad H_i \triangleq CA^{i-1}B \quad \text{for } i \geq 1 \tag{42.25}$$

Hence we need to develop a set of regression equations relating the true and modified Markov parameters.

First, from (42.16), recall that $\tilde{B}$ is partitioned into a part which multiplies the input and a part which multiplies the output. Hence we partition the modified Markov parameters the same way, that is, for $i \geq 1$, we let

$$\tilde{H}_i \triangleq C\tilde{A}^{i-1}\tilde{B} = C\tilde{A}^{i-1}\begin{bmatrix} B+GD, & -G \end{bmatrix} = \begin{bmatrix} \tilde{H}_{i,u}, & \tilde{H}_{i,y} \end{bmatrix} \tag{42.26}$$

Next, note that $H_0 = \tilde{H}_0 = D$. Hence

$$\tilde{H}_{1,u} + \tilde{H}_{1,y}H_0 = C(B+GD) - CGH_0 = CB = H_1 \tag{42.27}$$

Furthermore, for all $i \geq 2$, we find that

$$\begin{aligned} \tilde{H}_{i,u} + \tilde{H}_{i,y}H_0 &= C\tilde{A}^{i-1}B \\ &= C(A+GC)^{i-1}B \\ &= C(A+GC)^{i-2}AB + C(A+GC)^{i-2}GCB \\ &= CA^{i-1}B + \sum_{j=1}^{i-1} C(A+GC)^{i-j-1}GCA^{j-1}B \\ &= H_i - \sum_{j=1}^{i-1} \tilde{H}_{i-j,y}H_j \end{aligned} \tag{42.28}$$

and therefore the i^{th} Markov parameter of (42.1) is given by

$$H_i = \tilde{H}_{i,u} + \sum_{j=0}^{i-1} \tilde{H}_{i-j,y} H_j \tag{42.29}$$

Finally, let $\hat{\tilde{H}}_i$ denote the estimates of the modified Markov parameters given by (42.21), that is, let $\hat{\mathcal{H}}_n(\tilde{A}, \tilde{B}, C, D)$ be partitioned as

$$\hat{\mathcal{H}}_n(\tilde{A}, \tilde{B}, C, D) \triangleq \left[\begin{array}{cccccc} \hat{\tilde{H}}_0 & \hat{\tilde{H}}_{1,u} & \hat{\tilde{H}}_{1,y} & \cdots & \hat{\tilde{H}}_{n,u} & \hat{\tilde{H}}_{n,y} \end{array} \right] \tag{42.30}$$

Then using the recursion equation (42.29), we find that estimates of the true Markov parameters are given by

$$\hat{H}_i = \hat{\tilde{H}}_{i,u} + \sum_{j=0}^{i-1} \hat{\tilde{H}}_{i-j,y} \hat{H}_j \tag{42.31}$$

Furthermore, using the eigensystem realization algorithm (see Section **??**), we obtain estimates of the system matrices (A, B, C, D).

42.4 Summary

(1) Compute the least-squares estimate (42.21), partitioning the estimate into the Markov parameters which multiply the input and output, that is,

$$\hat{\mathcal{H}}_n(\tilde{A}, \tilde{B}, C, D) \triangleq \left[\begin{array}{cccccc} \hat{\tilde{H}}_0 & \hat{\tilde{H}}_{1,u} & \hat{\tilde{H}}_{1,y} & \cdots & \hat{\tilde{H}}_{n,u} & \hat{\tilde{H}}_{n,y} \end{array} \right] \tag{42.32}$$

(2) Compute the estimates $\hat{H}_0, \ldots, \hat{H}_n$ of the true system (42.1) using the recursion (42.31) with the estimates $\hat{\tilde{H}}_0$, $\hat{\tilde{H}}_{i,u}$, and $\hat{\tilde{H}}_{i,y}$,.

(3) Estimate the system matrices A, B, C, and D from the Markov parameter estimates $\hat{H}_0, \ldots, \hat{H}_n$ using the eigensystem realization algorithm (see Section **??**).

42.5 Equivalence to Normal Least-Squares

$$\begin{aligned} y(k) = v(k) &+ Du(k) \\ &+ \beta_{n-1}u(k-1) + \cdots + \beta_0 u(k-n) \\ &- \alpha_{n-1}y(k-1) - \cdots - \alpha_0 y(k-n) \\ &+ \gamma_{n-1}w(k-1) + \cdots + \gamma_0 w(k-n) \\ &+ \alpha_{n-1}v(k-1) + \cdots + \alpha_0 v(k-n) \end{aligned} \tag{42.33}$$

where, for all $i = 0, \ldots, n-1$,

$$\beta_i \triangleq C(A+GC)^{n-1-i}(B+GD) \tag{42.34}$$

$$\alpha_i \triangleq C(A+GC)^{n-1-i}M \tag{42.35}$$

$$\gamma_i \triangleq C(A+GC)^{n-1-i} \tag{42.36}$$

that is,

$$\alpha(\mathbf{q})y(k) = \beta(\mathbf{q})u(k) + \alpha(\mathbf{q})v(k) + \gamma(\mathbf{q})w(k) \tag{42.37}$$

42.6 Error Analysis

SECTION 43

(N4SID)

Consider the time-varying, discrete-time linear system

$$\begin{aligned} x(k+1) &= A(k)x(k) + B(k)u(k) + w(k) \\ y(k) &= C(k)x(k) + D(k)u(k) + v(k) \end{aligned} \tag{43.1}$$

where

$$\begin{aligned} A(k+1) &= f_A\Big(A(k)\Big), & B(k+1) &= f_B\Big(B(k)\Big) \\ C(k+1) &= f_C\Big(C(k)\Big), & D(k+1) &= f_D\Big(D(k)\Big) \end{aligned} \tag{43.2}$$

and

- $x \in \mathbb{R}^n$ denotes an unknown state vector.
- $u \in \mathbb{R}^m$ denotes a known (or measured) input.
- $y \in \mathbb{R}^p$ denotes a known (or measured) output.
- $w \in \mathbb{R}^n$ denotes an unknown process noise vector.
- $v \in \mathbb{R}^p$ denotes an unknown measurement noise vector.
- $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times m}$, $C \in \mathbb{R}^{p \times n}$, and $D \in \mathbb{R}^{p \times m}$ are unknown.
- f_A, f_B, f_C, and f_D are known.

Remark 43.1 The difference between this scenario and the standard discrete-time Kalman filter scenario (see Section 38) is that the system matrices (A, B, C, D) are now considered to be unknown.

Remark 43.2 If the system 43.1 is time-invariant, then for all $k \in \mathbb{Z}$,

$$\begin{aligned} A(k+1) &= A(k), & B(k+1) &= B(k) \\ C(k+1) &= C(k), & D(k+1) &= D(k) \end{aligned} \tag{43.3}$$

Next, consider the augmented matrices

$$f\Big(\tilde{x}(k), u(k)\Big) \triangleq \begin{bmatrix} A(k)x(k) \\ \texttt{vec}\Big(f_A\big(A(k)\big)\Big) \\ \texttt{vec}\Big(f_B\big(B(k)\big)\Big) \\ \texttt{vec}\Big(f_C\big(C(k)\big)\Big) \\ \texttt{vec}\Big(f_D\big(D(k)\big)\Big) \end{bmatrix}, \quad \tilde{x}(k) \triangleq \begin{bmatrix} x(k) \\ \texttt{vec}\Big(A(k)\Big) \\ \texttt{vec}\Big(B(k)\Big) \\ \texttt{vec}\Big(C(k)\Big) \\ \texttt{vec}\Big(D(k)\Big) \end{bmatrix} \tag{43.4}$$

$$h\Big(\tilde{x}(k), u(k)\Big) \triangleq C(k)x(k) + D(k)u(k)$$

where `vec` denotes the vectorization operator (see Appendix F). Then

$$\begin{aligned} \tilde{x}(k+1) &= f\Big(\tilde{x}(k), u(k)\Big) + w(k) \\ y(k) &= h\Big(\tilde{x}(k), u(k)\Big) + v(k) \end{aligned} \tag{43.5}$$

and hence the augmented system (43.5) is of the form required by the discrete-time extended Kalman filter which we developed in Section 40. Specifically, using the discrete-time extended Kalman filter, we can estimate the augmented state $\tilde{x}(k)$, which contains the system matrices (A, B, C, D). Therefore we can think of the discrete-time extended Kalman filter applied to our augemented system (43.5) as a system identification technique for our original state-space model (43.1).

PART XIII

Multivariable Polynomial Least-Squares

SECTION 44

Gröbner Bases

First, some notation and definitions:

- A *monomial* e in $x_1, \ldots, x_n$ is a product of the form
$$e = x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n} \tag{44.1}$$
where $\alpha_1, \ldots, \alpha_n$ are nonnegative integers.

- A *polynomial* f in $x_1, \ldots, x_n$ is a finite linear combination of monomials in $x_1, \ldots, x_n$, that is,
$$f = \sum_{i=1}^{k} a_i e_i \tag{44.2}$$
where k is a finite positive integer, $a_1, \ldots, a_k$ are scalars, and $e_1, \ldots, e_k$ are monomials in $x_1, \ldots, x_n$.

- If f is a polynomial in a single variable, for instance, if f is a polynomial in x_1, then f is called a *univariate polynomial.*

Example 44.1 Some polynomials in $x_1, \ldots, x_n$ are

$$f_1(x_1, \ldots, x_n) = 5 \tag{44.3}$$
$$f_2(x_1, \ldots, x_n) = x_2 \tag{44.4}$$
$$f_3(x_1, \ldots, x_n) = 7.5x_1 \tag{44.5}$$
$$f_4(x_1, \ldots, x_n) = -2x_2x_3 \tag{44.6}$$
$$f_5(x_1, \ldots, x_n) = 5.5x_2x_3 + x_1 + 1 - 0.2x_3^{10}x_2 \tag{44.7}$$

□

Definition 44.1 A Gröbner basis for a set of polynomials $f_1, \ldots, f_\ell$ in $x_1, \ldots, x_n$ is the set of polynomials $g_1, \ldots, g_m$ such that

$$\begin{gathered} f_1(x_1, \ldots, x_n) = \cdots = f_\ell(x_1, \ldots, x_n) = 0 \\ \text{if and only if} \\ g_1(x_1, \ldots, x_n) = \cdots = g_m(x_1, \ldots, x_n) = 0 \end{gathered} \tag{44.8}$$

Furthermore, there exists integers $i_1, i_2, i_3 \ldots, i_n$ such that

- g_{i_1} is a polynomial in x_1
- g_{i_2} is a polynomial in x_1 and x_2
- g_{i_3} is a polynomial in x_1, x_2, and x_3
- $\cdots$

□

Of course there are many other properties and details that I am leaving out, but that is the essence of the matter.

44.1 Implications

The main implication of the Gröbner basis is that we can solve g_{i_1} for all of the solutions x_1 using a standard univariate polynomial root solver. Once we have all of these solutions, we can plug them into g_{i_2}, one-by-one to obtain all of the solutions x_2 again by using a standard univariate polynomial root solver. Continuing in this manner, we can compute all of the solutions of our original polynomials $f_1, \ldots, f_n$.

44.2 Remarks

- The Gröbner basis can be computed using multivariate polynomial division algorithms, which can be computed in a very structured way.
- Generally, computing the Gröbner basis is done using symbolic solvers since the algorithm is based on polynomial division. This can sometimes lead to problems with finite precision numbers, however, there are also methods (like a paper of mine) which develops algorithms for this case.
- There does not exist (as far as I know) a method for determining the number of Gröbner basis polynomials *a priori*. Hence, for pathological cases, $m \gg n$, which means that for these cases, it may take a very long time to compute the Gr¨bner basis.

44.3 Special Cases: An Infinite Number of Solutions

There can exist an infinite number of solutions to the set of original polynomials $f_1, \ldots, f_n$. For instance, if our set is the single polynomial

$$f_1(x_1, x_2) = x_1 x_2 \tag{44.9}$$

then there exist an infinite number of solutions if x_1 and x_2 are allowed to be any real numbers. Similarly, the set of polynomials

$$f_1(x, y, z) = x^2 + y^2 + z^2 - 1 \tag{44.10}$$

$$f_2(x, y, z) = xyz - 1 \tag{44.11}$$

has the Gröbner basis

$$g_1(z) = 0 \tag{44.12}$$

$$g_2(y, z) = y^4 z^2 + y^2 z^4 - y^2 z^2 + 1 \tag{44.13}$$

$$g_3(x, y, z) = x + y^3 z + y z^3 - yz \tag{44.14}$$

Hence our polynomial g_1 in z, which is required to be univariate in z is 0, that is, all z's are permissible solutions. Thus there are an infinite number of solutions of f_1, f_2.

However, if you impose the additional polynomial which constrains all of the variables to be binary, there cannot be an infinite number of solutions, and hence you can ignore this case.

PART XIV

Appendices

SECTION A

Double Summations

$$\sum_{k=0}^{m}\sum_{j=0}^{k} a_{k,j} = \sum_{j=0}^{m}\sum_{k=j}^{m} a_{k,j} \tag{A.1}$$

$$\sum_{k=0}^{m}\sum_{j=k}^{m} a_{k,j} = \sum_{j=0}^{m}\sum_{k=0}^{j} a_{k,j} \tag{A.2}$$

SECTION B

Matrix Inversion Lemma

The *matrix inversion lemma*, or *Woodbury matrix identity*, is one of the most useful identities for inverting sums and products of matrices. Several forms of the identity can be found in [8, Corollary 2.8.8], with the two most common forms given in the following lemma:

Lemma B.1 Let $A \in \mathbb{C}^{n\times n}$, $B \in \mathbb{C}^{n\times m}$, $C \in \mathbb{C}^{m\times n}$, and $D \in \mathbb{C}^{m\times m}$. If A, D, and $D - CA^{-1}B$ are invertible, then $A - BD^{-1}C$ is invertible. Specifically,

$$\big(A - BD^{-1}C\big)^{-1} = A^{-1} + A^{-1}B\big(D - CA^{-1}B\big)^{-1}CA^{-1} \tag{B.1}$$

$$C\big(A - BD^{-1}C\big)^{-1}A = D\big(D - CA^{-1}B\big)^{-1}C \tag{B.2}$$

SECTION C

Block Matrix Inverse

Fact C.1 Let $A, E \in \mathbb{R}^{m\times m}$, $B, F \in \mathbb{R}^{m\times n}$, $C, G \in \mathbb{R}^{n\times m}$, and $D, H \in \mathbb{R}^{n\times n}$, where A and D are invertible. Then

- $\left(A - BD^{-1}C\right)^{-1}$ is invertible if and only if $\left(D - CA^{-1}B\right)^{-1}$ is invertible.
- If $\left(A - BD^{-1}C\right)^{-1}$ is invertible, then $\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1}$ is given by

$$\begin{bmatrix} E & F \\ G & H \end{bmatrix} \triangleq \begin{bmatrix} \left(A - BD^{-1}C\right)^{-1} & -A^{-1}B\left(D - CA^{-1}B\right)^{-1} \\ -D^{-1}C\left(A - BD^{-1}C\right)^{-1} & \left(D - CA^{-1}B\right)^{-1} \end{bmatrix} \tag{C.1}$$

Proof The first statement follows directly from the matrix inversion lemma, Lemma B.1.

Next, note that (C.1) is the inverse of $\begin{bmatrix} A & B \\ C & D \end{bmatrix}$ if and only if

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} E & F \\ G & H \end{bmatrix} = \begin{bmatrix} I_m & 0_{m\times n} \\ 0_{n\times m} & I_n \end{bmatrix}$$

that is,

$$\begin{aligned} AE + BG &= I_m, & AF + BH &= 0_{m\times n} \\ CE + DG &= 0_{n\times m}, & CF + DH &= I_n \end{aligned}$$

Specifically, since A and D are invertible, it follows that

$$\begin{aligned} E &= A^{-1}\left(I_m - BG\right), & F &= -A^{-1}BH \\ H &= D^{-1}\left(I_n - CF\right), & G &= -D^{-1}CE \end{aligned}$$

Finally, substituting F and G into E and H, respectively, we find that

$$\begin{aligned} E = A^{-1}\left(I_n + BD^{-1}CE\right) \quad &\Longrightarrow \quad E = \left(I_m - A^{-1}BD^{-1}C\right)^{-1}A^{-1} \\ & = \left(A - BD^{-1}C\right)^{-1} \\ H = D^{-1}\left(I_m + CA^{-1}BH\right) \quad &\Longrightarrow \quad H = \left(I_n - D^{-1}CA^{-1}B\right)^{-1}D^{-1} \\ & = \left(D - CA^{-1}B\right)^{-1} \end{aligned}$$

Therefore substituting E and H back into F and G, we find (C.1). □

SECTION D

Nullspaces

Fact D.1 Let $A \in \mathbb{R}^{m \times n}$ have full row rank. Then

$$V \triangleq I_n - A^T \left(AA^T\right)^{-1} A \tag{D.1}$$

is an overdetermined basis for the nullspace of A.

Proof First, note that

$$AV = A - AA^T \left(AA^T\right)^{-1} A = A - A = 0_{n \times n} \tag{D.2}$$

Furthermore, note that from the structure of the problem we must have that

$$\texttt{rank}\big[A\big] + \texttt{rank}\big[V\big] \leq n$$

which holds with equality if V is a basis for the nullspace of A.

Next, from the structure of V, note that

$$\texttt{rank}\big[V\big] \geq n - \texttt{rank}\big[A\big]$$

Hence combining, we find that

$$\texttt{rank}\big[A\big] + \texttt{rank}\big[V\big] = n$$

that is, V is an overdetermined basis for the nullspace of A, where we say that V is an overdetermined basis since it does not have full column rank. □

SECTION E

The Matrix Trace

Definition E.1 Let $A \in \mathbb{C}^{n \times n}$ be given by

$$A = \begin{bmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{n,1} & \cdots & a_{n,n} \end{bmatrix} \tag{E.1}$$

Then the *trace of A* is the sum of all of the diagonal elements of A. Specifically, we write

$$\texttt{tr}[A] = \sum_{i=1}^{n} a_{i,i} \tag{E.2}$$

□

Fact E.1 If $A \in \mathbb{C}^{n \times n}$, then $\texttt{tr}[A^T] = \texttt{tr}[A]$.

Proof The diagonal elements of a square matrix are unchanged under transposition. □

Fact E.2 If $A \in \mathbb{C}^{n \times n}$, then

$$\texttt{tr}[A^H] = \texttt{tr}[\overline{A}] = \overline{\texttt{tr}[A]} \tag{E.3}$$

where $\overline{A}$ and A^H denote the complex conjugate and Hermitian, respectively, that is, $A^H = \overline{A}^T$.

Proof From Fact E.1, $\texttt{tr}[A^H] = \texttt{tr}[\overline{A}]$. Therefore, since the sum of conjugates is equal to the conjugate of the sum, we have that

$$\texttt{tr}[A^H] = \texttt{tr}[\overline{A}] = \sum_{i=1}^{n} \overline{a_{i,i}} = \overline{\left(\sum_{i=1}^{n} a_{i,i} \right)} = \overline{\texttt{tr}[A]} \tag{E.4}$$

□

Fact E.3 If $A \in \mathbb{C}^{m \times n}$ and $B \in \mathbb{C}^{n \times m}$, then

$$\texttt{tr}[BA] = \sum_{k=1}^{n} \sum_{j=1}^{m} b_{k,j} a_{j,k} = \sum_{j=1}^{m} \sum_{k=1}^{n} a_{j,k} b_{k,j} = \texttt{tr}[AB] \tag{E.5}$$

Proof First, let A and B be given by

$$A = \begin{bmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{m,1} & \cdots & a_{m,n} \end{bmatrix}, \qquad B = \begin{bmatrix} b_{1,1} & \cdots & b_{1,m} \\ \vdots & \ddots & \vdots \\ b_{n,1} & \cdots & b_{n,m} \end{bmatrix}$$

Then the j^{th} diagonal element of AB, and the k^{th} diagonal element of BA are given by

$$(AB)_j = \sum_{k=1}^{n} a_{j,k} b_{k,j}, \qquad (BA)_k = \sum_{j=1}^{m} b_{k,j} a_{j,k}$$

Hence the trace of AB and BA are given by

$$\begin{aligned} \texttt{tr}[AB] &= \sum_{j=1}^{m} (AB)_j = \sum_{j=1}^{m} \sum_{k=1}^{n} a_{j,k} b_{k,j} \\ \texttt{tr}[BA] &= \sum_{k=1}^{n} (BA)_k = \sum_{k=1}^{n} \sum_{j=1}^{m} b_{k,j} a_{j,k} \end{aligned}$$

Therefore, switching the summation order in $\texttt{tr}[BA]$, we find that $\texttt{tr}[BA] = \texttt{tr}[AB]$, that is, (E.5). □

Remark E.1 To show (E.5), you need to show that the sum of the diagonal elements of AB and BA are equal. This is different than showing that the i^{th} diagonal element of AB is equal to the i^{th} diagonal element of BA. Specifically, since A and B are not necessarily square matrices, AB and BA may have different dimensions, and hence they might have a different number of diagonal elements. Therefore comparing their diagonal elements does not even make sense in many cases. □

E.1 Trace Derivatives

Fact E.4 Let $A \in \mathbb{C}^{m\times n}$ and $X \triangleq \begin{bmatrix} x_{1,1} & \cdots & x_{1,m} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,m} \end{bmatrix} \in \mathbb{C}^{n\times m}$. Then

$$\frac{\partial \mathtt{tr}\left[AX\right]}{\partial X} \triangleq \begin{bmatrix} \dfrac{\partial \mathtt{tr}\left[AX\right]}{\partial x_{1,1}} & \cdots & \dfrac{\partial \mathtt{tr}\left[AX\right]}{\partial x_{1,m}} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial \mathtt{tr}\left[AX\right]}{\partial x_{n,1}} & \cdots & \dfrac{\partial \mathtt{tr}\left[AX\right]}{\partial x_{n,m}} \end{bmatrix}^T = A \tag{E.6}$$

Proof Let $k \in [1, n]$ and $j \in [1, m]$. Then

$$\frac{\partial \mathtt{tr}\left[AX\right]}{\partial x_{k,j}} = \frac{\partial\left(\sum\limits_{j=1}^{m}\sum\limits_{k=1}^{n} a_{j,k}x_{k,j}\right)}{\partial x_{k,j}} = a_{j,k}$$

and hence

$$\frac{\partial \mathtt{tr}\left[AX\right]}{\partial X} \triangleq \begin{bmatrix} a_{1,1} & \cdots & a_{m,1} \\ \vdots & \ddots & \vdots \\ a_{1,n} & \cdots & a_{m,n} \end{bmatrix}^T = A$$

□

Remark E.2 The matrix in (E.6) is, by convention, transposed. When A and X are complex matrices, this remains the transpose, and not the Hermitian matrix, despite what your experience may tell you. □

Remark E.3 We use the convention that the derivative is transposed, as in (E.6). The reason for this convention is that, in the scalar case, you will often see the first-order approximation

$$f(x) \approx \frac{\partial f(x)}{\partial x}\Delta x \tag{E.7}$$

where $x \in \mathbb{C}$, $f : \mathbb{C} \to \mathbb{C}$, and $\Delta x \in \mathbb{C}$ denotes a perturbation to x. Now, if we let x be a vector, that is, if we let $x \in \mathbb{C}^{n\times 1}$, then we still want to be able to make statements like (E.7). The way to accomplish this is by defining the derivative to have the same dimension as the transpose of x, that is,

$$\frac{\partial f(x)}{\partial x} \triangleq \left[\begin{array}{ccc} \frac{\partial f(x)}{\partial x_1} & \cdots & \frac{\partial f(x)}{\partial x_n} \end{array} \right] \in \mathbb{C}^{1\times n} \tag{E.8}$$

so that the first order approximation can be compactly written as

$$f(x) \approx \frac{\partial f(x)}{\partial x}\Delta x = \left[\begin{array}{ccc} \frac{\partial f(x)}{\partial x_1} & \cdots & \frac{\partial f(x)}{\partial x_n} \end{array} \right] \left[\begin{array}{c} \Delta x_1 \\ \vdots \\ \Delta x_n \end{array} \right] = \sum_{i=1}^{n} \frac{\partial f(x)}{\partial x_i}\Delta x_i$$

In the matrix case, that is, when $x \in \mathbb{C}^{n\times m}$, the result is not as nice since $\dfrac{\partial f(x)}{\partial x}\Delta x$ evaluates to a matrix, but at the very least, we can still multiply $\partial f(x)/\partial x$ and Δx without introducing a bunch of transposes. □

Fact E.5 Let $A \in \mathbb{C}^{m\times n}$, $X \in \mathbb{C}^{n\times p}$, and $B \in \mathbb{R}^{p\times m}$. Then

$$\frac{\partial \mathtt{tr}\big[AXB\big]}{\partial X} = \frac{\partial \mathtt{tr}\big[BAX\big]}{\partial X} = \frac{\partial \mathtt{tr}\big[XBA\big]}{\partial X} = BA \tag{E.9}$$

Proof From Fact E.3, it follows that

$$\mathtt{tr}\big[AXB\big] = \mathtt{tr}\big[BAX\big] = \mathtt{tr}\big[XBA\big] \tag{E.10}$$

Hence using Fact E.4, we find (E.9). □

Fact E.6 Let $A, X \in \mathbb{C}^{m\times n}$. Then

$$\frac{\partial \mathtt{tr}\big[AX^T\big]}{\partial X} = A^T \tag{E.11}$$

Proof From Fact E.1, it follows that $\mathtt{tr}\big[AX^T\big] = \mathtt{tr}\big[XA^T\big]$. Therefore, from Fact E.5, we find (E.11). □

Fact E.7 Let $A, B \in \mathbb{C}^{p\times n}$ and $X \in \mathbb{C}^{n\times p}$. Then

$$\frac{\partial \mathtt{tr}\big[AXBX\big]}{\partial X} = BXA + AXB \tag{E.12}$$

Proof Let $Y \triangleq BX$ and $Z \triangleq AXB$. Then from the product rule for derivatives, we have that

$$\frac{\partial \mathtt{tr}\left[AXBX\right]}{\partial X} = \frac{\partial \mathtt{tr}\left[AXY\right]}{\partial X} + \frac{\partial \mathtt{tr}\left[ZX\right]}{\partial X} \tag{E.13}$$

Hence, from Fact E.5, we find that

$$\frac{\partial \mathtt{tr}\left[AXBX\right]}{\partial X} = YA + Z \tag{E.14}$$

that is, (E.12). □

Fact E.8 Let $A \in \mathbb{C}^{n \times n}$, $X \in \mathbb{C}^{n \times p}$, and $B \in \mathbb{C}^{p \times p}$. Then

$$\frac{\partial \mathtt{tr}\left[AXBX^T\right]}{\partial X} = BX^TA + B^TX^TA^T \tag{E.15}$$

Proof Let $Y \triangleq BX^T$ and $Z \triangleq AXB$. Then from the product rule for derivatives, we have that

$$\frac{\partial \mathtt{tr}\left[AXBX^T\right]}{\partial X} = \frac{\partial \mathtt{tr}\left[AXY\right]}{\partial X} + \frac{\partial \mathtt{tr}\left[ZX^T\right]}{\partial X} \tag{E.16}$$

Hence, from Fact E.1 and Fact E.5, we find that

$$\frac{\partial \mathtt{tr}\left[AXBX^T\right]}{\partial X} = \frac{\partial \mathtt{tr}\left[AXY\right]}{\partial X} + \frac{\partial \mathtt{tr}\left[XZ^T\right]}{\partial X} = YA + Z^T \tag{E.17}$$

that is, (E.15). □

Fact E.9 Let $X \in \mathbb{C}^{n \times p}$. Then

$$\frac{\partial \mathtt{tr}\left[XX^H\right]}{\partial X} = X^HA + X^TA^T \tag{E.18}$$

Proof Let $x_{i,j}$ denote the $(i,j)^{th}$ element of X. Then

$$\frac{\partial \mathtt{tr}\left[XX^H\right]}{\partial X} = \frac{\partial \left(\sum\limits_{i=1}^{n} \sum\limits_{j=1}^{p} x_{i,j} x_{i,j}^H \right)}{\partial X} = \frac{\partial \left(\sum\limits_{i=1}^{n} \sum\limits_{j=1}^{p} |x_{i,j}|^2 \right)}{\partial X} \tag{E.19}$$

□

E.2 Derivatives with respect to Complex Numbers

Cauchy Riemann. Show that the complex conjugate is not differentiable. We need the alternate definition of a complex derivative given on page 24 of the matrix cookbook.

E.3 The Least-Squares Solution

Fact E.10 Let $A \in \mathbb{C}^{n\times n}$, $X \in \mathbb{C}^{n\times p}$, and $B \in \mathbb{C}^{p\times p}$. Then

$$\frac{\partial \mathtt{tr}\left[AXBB^H X^H\right]}{\partial X} = BB^H X^H A + B^T X^T A^T \tag{E.20}$$

Proof TODO □

Fact E.11 Trace properties. Note that

$$\mathtt{tr}\left[(\Theta A - B)^H (\Theta A - B)\right] = \mathtt{tr}\left[A^H \Theta^H \Theta A - A^H \Theta^H B - B^H \Theta A + B^H B\right] \tag{E.21}$$

Hence when A, B, and Θ are real, we have that

$$\mathtt{tr}\left[A^H \Theta^H B\right] = \mathtt{tr}\left[A^T \Theta^T B\right] = \mathtt{tr}\left[B^H \Theta A\right] \tag{E.22}$$

and therefore

$$\mathtt{tr}\left[(\Theta A - B)^H (\Theta A - B)\right] = \mathtt{tr}\left[A^H \Theta^H \Theta A - 2B^H \Theta A + B^H B\right] \tag{E.23}$$

E.4 Other Identities

Fact E.12 Let $A \in \mathbb{C}^{n\times n}$, and let $\lambda_1, \ldots, \lambda_n$ denote the eigenvalues of A. Then

(E.24)

SECTION F

Kronecker Product and Vec Operator

Definition F.1 Let $A \in \mathbb{C}^{m\times n}$ and $B \in \mathbb{C}^{p\times q}$, where

$$A \triangleq \begin{bmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & & \vdots \\ a_{m,1} & \cdots & a_{m,n} \end{bmatrix} \tag{F.1}$$

Then the *Kronecker product of* A *and* B, denoted by $A \otimes B$, is given by

$$A \otimes B \triangleq \begin{bmatrix} a_{1,1}B & \cdots & a_{1,n}B \\ \vdots & & \vdots \\ a_{m,1}B & \cdots & a_{m,n}B \end{bmatrix} \in \mathbb{C}^{mp\times nq} \tag{F.2}$$

Furthermore, letting $\texttt{col}_i(A)$ denote the i^{th} column of A, we have that the *vec of* A is the vector

$$\texttt{vec}(A) \triangleq \begin{bmatrix} \texttt{col}_1(A) \\ \vdots \\ \texttt{col}_n(A) \end{bmatrix} \in \mathbb{C}^{mn\times 1} \tag{F.3}$$

□

Remark F.1 In general, the Kronecker product does not commute, that is, in general, $A \otimes B \neq B \otimes A$. □

Remark F.2 Although rarely mentioned, the Kronecker product is the standard operation used for multiplying scalars with matrices. For instance, if $\alpha \in \mathbb{C}$, and $A \in \mathbb{C}^{m\times n}$ is given by (F.1), then

$$A \otimes \alpha = \begin{bmatrix} \alpha a_{1,1} & \cdots & \alpha a_{1,n} \\ \vdots & & \vdots \\ \alpha a_{m,1} & \cdots & \alpha a_{m,n} \end{bmatrix} = \alpha \otimes A = \alpha A \tag{F.4}$$

□

From [8], we also have the following useful facts about the Kronecker product

Fact F.1 If $A \in \mathbb{C}^{m\times n}$ and $B, C \in \mathbb{C}^{p\times q}$, then

$$(A \otimes B) \otimes C = A \otimes (B \otimes C) \tag{F.5}$$

$$A \otimes (B + C) = A \otimes B + A \otimes C \tag{F.6}$$

$$(B + C) \otimes A = B \otimes A + C \otimes A \tag{F.7}$$

$$(A \otimes B)^T = A^T \otimes B^T \tag{F.8}$$

$$\texttt{rank}(A \otimes B) = \texttt{rank}(A)\texttt{rank}(B) \tag{F.9}$$

If $A \in \mathbb{C}^{m\times n}$, $B \in \mathbb{C}^{p\times q}$, $C \in \mathbb{C}^{n\times k}$, and $D \in \mathbb{C}^{q\times \ell}$, then

$$(A \otimes B)(C \otimes D) = AC \otimes BD \tag{F.10}$$

Finally, if A and B are invertible, then

$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1} \tag{F.11}$$

The following fact relates the vec operator and the matrix trace (see Appendix E):

Fact F.2

$$\texttt{tr}\big(AC\big) = \texttt{vec}\big(A^T\big)^T \texttt{vec}\big(C\big) = \texttt{vec}\big(C^T\big)^T \texttt{vec}\big(A\big) \tag{F.12}$$

The following fact relates the vec operator and the Kronecker product:

Fact F.3

$$\texttt{vec}\big(ABC\big) = \big(C^T \otimes A\big)\, \texttt{vec}\big(B\big) \tag{F.13}$$

SECTION G

Toeplitz Matrices

Fact G.1 If $V \in \mathbb{R}^{\ell\times k}$ is nonzero, then

$$\mathtt{rank}\left[\begin{array}{c|c} V & 0_{1\times k} \\ 0_{1\times k} & V \end{array}\right] > \mathtt{rank}\big[V\big] \tag{G.1}$$

Proof Since V is nonzero, there exists at least one nonzero row of V. Specifically, letting $v_j \in \mathbb{R}^{1\times k}$ denote the last nonzero row of V, we have that

$$\left[\begin{array}{c|c} V & 0_{1\times k} \\ 0_{1\times k} & V \end{array}\right] = \left[\begin{array}{c|c} \tilde{V} & 0_{1\times k} \\ v_\ell & \tilde{V} \\ \hline 0_{1\times k} & v_j \\ 0_{(\ell-j)\times k} & 0_{(\ell-j)\times k} \end{array}\right]$$

Hence the last k columns of $\left[\begin{array}{c|c} V & 0_{1\times k} \\ 0_{1\times k} & V \end{array}\right]$ are not in the range of the first k columns, that is, (G.1). $\square$

Fact G.2 Let

$$\tilde{\Phi} \triangleq \begin{bmatrix} A_0 & \tilde{A} & A_\ell \\ B_0 & \tilde{B} & B_\ell \end{bmatrix} \in \mathbb{R}^{(m+n)\times(x+2)} \tag{G.2}$$

where $A_0, A_\ell \in \mathbb{R}^m$, $B_0, B_\ell \in \mathbb{R}^n$, $\tilde{A} \in \mathbb{R}^{m\times x}$, and $\tilde{B} \in \mathbb{R}^{n\times x}$. Furthermore, let $V \in \mathbb{R}^{(x+1)\times k}$ be a basis for the nullspace of $\begin{bmatrix} A_0, & \tilde{A} \\ B_0, & \tilde{B} \end{bmatrix}$, where

$$\mathtt{rank}\begin{bmatrix} A_0, & \tilde{A} \\ B_0, & \tilde{B} \end{bmatrix} = \mathtt{rank}\begin{bmatrix} A_0, & \tilde{A} \end{bmatrix} + \mathtt{rank}\begin{bmatrix} B_0, & \tilde{B} \end{bmatrix} \tag{G.3}$$

If $\begin{bmatrix} \tilde{B} & B_\ell \end{bmatrix} V = 0$ and $\mathtt{rank}\begin{bmatrix} B_0, & \tilde{B}, & B_\ell \end{bmatrix} = \mathtt{rank}\begin{bmatrix} B_0, & \tilde{B} \end{bmatrix}$, then

then there exists $W \in \mathbb{R}^{2k\times q}$ such that $\begin{bmatrix} V & 0_{1\times k} \\ 0_{1\times k} & V \end{bmatrix} W$ is a basis for the nullspace of $\tilde{\Phi}$.

Proof Since $V \in \mathbb{R}^{(x+1)\times k}$ is a basis for the nullspace of $\left[\begin{array}{cc} A_0, & \tilde{A} \\ B_0, & \tilde{B} \end{array}\right]$,

$$\texttt{rank}\left[\begin{array}{cc} A_0, & \tilde{A} \\ B_0, & \tilde{B} \end{array}\right] + \texttt{rank}\big[V\big] = x+1$$

Therefore, from (G.3), we have that

$$\texttt{rank}\left[\begin{array}{cc} B_0, & \tilde{B} \end{array}\right] + \texttt{rank}\big[V\big] = x+1-\texttt{rank}\left[\begin{array}{cc} A_0, & \tilde{A} \end{array}\right]$$

Hence, although $\left[\begin{array}{cc} B_0, & \tilde{B} \end{array}\right] V = 0$, V is not a complete basis for the nullspace of $\left[\begin{array}{cc} B_0, & \tilde{B} \end{array}\right]$. Specifically, letting $p \triangleq \texttt{rank}\left[\begin{array}{cc} A_0, & \tilde{A} \end{array}\right]$, it follows that there exists a $W \in \mathbb{R}^{(x+1)\times p}$ with full column rank, which is linearly independent of V, and which satisfies $\left[\begin{array}{cc} B_0, & \tilde{B} \end{array}\right] W = 0$. Specifically, let's chose W such that

$$\left[\begin{array}{cc} A_0, & \tilde{A} \end{array}\right] W = I$$

Next, let

$$\tilde{V} \triangleq \left[\begin{array}{c|c} \begin{array}{c} V, \\ 0_{1\times k} \end{array} & \left(I_{x+2} - \left[\begin{array}{c} W \\ 0_{1\times p} \end{array}\right]\left[\begin{array}{ccc} A_0, & \tilde{A}, & A_\ell \end{array}\right]\right)\left[\begin{array}{c} 0_{1\times k} \\ V \end{array}\right] \end{array}\right]$$

Then from Fact G.1, $\texttt{rank}\big[\tilde{V}\big] > \texttt{rank}\big[V\big]$. Furthermore, $\tilde{\Phi}\tilde{V} = 0$. Hence $\tilde{V}$ is a basis for the nullspace of $\tilde{\Phi}$ (although it is probably overdetermined). □

Let $N > q$ and $p \triangleq \texttt{floor}([q+1]/\ell)$. Then

$$\Phi_{q,N} = \left[\begin{array}{ccc} u(q+1) & \cdots & u(N) \\ \vdots & & \vdots \\ u(p\ell+1) & \cdots & u(N-q+p\ell) \\ \hline u(p\ell) & \cdots & u(N-q+p\ell-1) \\ \vdots & & \vdots \\ u([p-1]\ell+1) & \cdots & u(N-q+[p-1]\ell) \\ \hline u([p-1]\ell) & \cdots & u(N-q+[p-1]\ell-1) \\ \vdots & & \vdots \\ u([p-2]\ell+1) & \cdots & u(N-q+[p-2]\ell) \\ \hline \vdots & & \vdots \\ \hline u(\ell) & \cdots & u(N-q+\ell-1) \\ \vdots & & \vdots \\ u(1) & \cdots & u(N-q) \end{array}\right] \tag{G.4}$$

and hence

$$\Phi_{q,N} = \left[\begin{array}{c} \left[\begin{array}{cc} I_{m([q+1]-p\ell)}, & 0_{m([q+1]-p\ell)\times m([p+1]\ell-[q+1])} \end{array}\right] \Phi_{\ell-1,N} \left[\begin{array}{c} 0_{(q-\ell+1)\times(N-q)} \\ I_{(N-q)} \end{array}\right] \\ \hline \Phi_{\ell-1,N} \left[\begin{array}{c} 0_{[p-1]\ell\times(N-q)} \\ I_{(N-q)} \\ 0_{([q+1]-p\ell)\times(N-q)} \end{array}\right] \\ \hline \Phi_{\ell-1,N} \left[\begin{array}{c} 0_{[p-2]\ell\times(N-q)} \\ I_{(N-q)} \\ 0_{([q+1]-[p-1]\ell)\times(N-q)} \end{array}\right] \\ \hline \vdots \\ \hline \Phi_{\ell-1,N} \left[\begin{array}{c} I_{(N-q)} \\ 0_{([q+1]-\ell)\times(N-q)} \end{array}\right] \end{array}\right. \tag{G.5}$$

Fact G.3

Proof Next, note that $\Phi_{\ell-1,N-1}$ has full row rank. Hence letting V denote a basis for the nullspace of $\Phi_{\ell-1,N-1}$, it follows that $V \in \mathbb{R}^{(N-\ell)\times(N-\ell-m\ell)}$, where V has full column rank, and for $k = 1, \ldots, \ell$,

$$\begin{bmatrix} u(k) & \cdots & u(N-\ell-1+k) \end{bmatrix} V = 0_{m\times(N-\ell-m\ell)}$$

Furthermore, since

$$\begin{aligned} \Phi_{\ell-1,N} &= \begin{bmatrix} u(\ell) & u(\ell+1) & \cdots & u(N-1) & u(N) \\ u(\ell-1) & u(\ell) & \cdots & u(N-2) & u(N-1) \\ \vdots & \vdots & & \vdots & \vdots \\ u(1) & u(2) & \cdots & u(N-\ell) & u(N-\ell+1) \end{bmatrix} \\ &= \left[\begin{array}{c} u(\ell) \quad \cdots \quad u(N) \\ \hline \Phi_{\ell-2,N-1} \end{array}\right] \end{aligned}$$

it follows that

$$\Phi_{\ell-2,N-1} \left[\begin{array}{c|c} V & 0_{1\times(N-\ell-m\ell)} \\ 0_{1\times(N-\ell-m\ell)} & V \end{array}\right] = 0_{m(\ell-1)\times 2(N-\ell-m\ell)}$$

where, from Fact G.1, we have that

$$\texttt{rank}\left[\begin{array}{c|c} V & 0_{1\times(N-\ell-m\ell)} \\ 0_{1\times(N-\ell-m\ell)} & V \end{array}\right] > \texttt{rank}\big[V\big] = N - \ell - m\ell$$

Next, note that if a matrix U is in the nullspace of $\Phi_{\ell-2,N-1}$, that is, $\Phi_{\ell-2,N-1}U = 0$, then

$$\texttt{rank}\big[\Phi_{\ell-2,N-1}\big] + \texttt{rank}\big[U\big] \leq N - \ell + 1$$

where, since $\Phi_{\ell-2,N-1}$ has full row rank,

$$\texttt{rank}\big[U\big] \leq N - \ell + 1 - m\ell + m$$

Furthermore, since $\Phi_{\ell-2,N-1}$ has full row rank, it follows that

$$\texttt{rank}\big[\Phi_{\ell-2,N-1}\big] + \texttt{rank}\left[\begin{array}{c|c} V & 0_{1\times(N-\ell-m\ell)} \\ 0_{1\times(N-\ell-m\ell)} & V \end{array}\right] > N - \ell - m$$

Next, comes the question... Is $\left[\begin{array}{c|c} V & 0_{1\times(N-\ell-m\ell)} \\ 0_{1\times(N-\ell-m\ell)} & V \end{array}\right]$ a basis for the nullspace of $\Phi_{\ell-2,N-1}$.

Note that $\Phi_{\ell-1,N}$ and $\Phi_{\ell-1,N-1}$, and $\Phi_{\ell-2,N-1}$ have full row rank. Hence their nullspaces are of dimension $N-\ell+1-m\ell$ and $N-\ell-m\ell$, respectively. Specifically, letting $V_{i,j}$ denote a basis for the nullspace of $\Phi_{i,j}$, then for all $k = 1\ldots,\ell$,

$$\left[\begin{array}{ccc} u(k) & \cdots & u(N-\ell+k) \end{array}\right] V_{\ell-1,N-1} = 0_{\times(N-\ell-m\ell+m)}$$

Next, note that $\Phi_{\ell-1,N-1}$ has full row rank, where

$$\Phi_{\ell-1,N-1} = \left[\begin{array}{ccc} u(\ell) & \cdots & u(N-1) \\ \hline & \Phi_{\ell-2,N-2} & \end{array}\right] \in \mathbb{R}^{m\ell\times(N-\ell)}$$

Hence letting V denote a basis for the nullspace of the final $m\ell-1$ rows of $\Phi_{\ell-1,N-1}$, it follows that $V \in \mathbb{R}^{(N-\ell)\times(N-\ell-m\ell+1)}$, where V has full column rank, and for $k = 1,\ldots,\ell-1$,

$$\left[\begin{array}{ccc} u(k) & \cdots & u(N-\ell-1+k) \end{array}\right] V = 0_{m\times(N-\ell-m\ell+1)}$$

Furthermore, the final $m-1$ rows of the product $\left[\begin{array}{ccc} u(\ell) & \cdots & u(N-1) \end{array}\right] V$ are also zero.

□

SECTION H

Polynomial Matrix Facts

Definition H.1 Consider the nonzero operator polynomial $C \in \mathbb{R}^{p\times m}[\mathbf{q}]$ given by

$$C(\mathbf{q}) \triangleq C_\ell \mathbf{q}^\ell + \cdots + C_1\mathbf{q} + C_0 \tag{H.1}$$

If $C_i \neq 0$ and $C_{i+1}, \ldots, C_\ell = 0$, then C_i is called the *leading coefficient* of $C(\mathbf{q})$. If $C_i \neq 0$ and $C_0, \ldots, C_{i-1} = 0$, then C_i is called the *trailing coefficient* of $C(\mathbf{q})$. □

Definition H.2 Let $C \in \mathbb{R}^{m\times m}[\mathbf{q}]$. Then $C(\mathbf{q})$ has full normal rank if

$$\mathtt{det}\big[C(\mathbf{q})\big] \neq 0 \tag{H.2}$$

□

Fact H.1 Let $C, E \in \mathbb{R}^{m\times m}[\mathbf{q}]$, where $C(\mathbf{q})$ has full normal rank. Then $E(\mathbf{q})C(\mathbf{q}) = 0_{m\times m}$ if and only if $E(\mathbf{q}) = 0_{m\times m}$.

Proof First, note that if $E(\mathbf{q}) = 0_{m\times m}$, then $E(\mathbf{q})C(\mathbf{q}) = 0_{m\times m}$.

Next, note that if $E(\mathbf{q})C(\mathbf{q}) = 0_{m\times m}$, then right-multiplying by the matrix adjugate of C, we find that

$$E(\mathbf{q})C(\mathbf{q})\mathtt{adj}\big[C(\mathbf{q})\big] = E(\mathbf{q})\mathtt{det}\big[C(\mathbf{q})\big] = 0_{m\times m}$$

where $\mathtt{det}\big[C(\mathbf{q})\big] \neq 0$ since $C(\mathbf{q})$ has full normal rank. Specifically, letting $\bar{E}$ and $\bar{\alpha}$ denote the leading coefficients of $E(\mathbf{q})$ and $\mathtt{det}\big[C(\mathbf{q})\big]$, respectively, we find that leading coefficient of the product $E(\mathbf{q})\mathtt{det}\big[C(\mathbf{q})\big]$ is $\bar{E}\bar{\alpha}$. However, since $E(\mathbf{q})\mathtt{det}\big[C(\mathbf{q})\big] = 0_{m\times m}$ and $C(\mathbf{q})$ has full normal rank, it follows that $\bar{E}\bar{\alpha} = 0_{m\times m}$ and $\bar{\alpha} \neq 0$. Thus $\bar{E} = 0_{m\times m}$, that is, the leading coefficient of $E(\mathbf{q})$ is zero. Hence $E(\mathbf{q}) = 0_{m\times m}$. □

Fact H.2 If $C \in \mathbb{R}^{m\times m}[\mathbf{q}]$ has full normal rank, then for all $i \geq 0$, $\mathcal{T}_i(C)$ has full row rank.

Proof Suppose that there exists an $i \geq 0$ for which $\mathcal{T}_i(C)$ does not have full row rank. Then there exists a nonzero vector $\tilde{E} \in \mathbb{R}^{m\times m(i+1)}$ in the left nullspace of $\mathcal{T}_i(C)$. Specifically, partitioning $\tilde{E}$ into $m \times m$ blocks, that is, $\tilde{E} \triangleq \left[\begin{array}{ccc} E_i & \cdots & E_0 \end{array} \right]$ and letting

$$E(\mathbf{q}) \triangleq E_i\mathbf{q}^i + \cdots + E_1\mathbf{q} + E_0$$

then from Fact 12.8 we have that

$$\tilde{E}\mathcal{T}_i(C) = 0 \quad \Longleftrightarrow \quad E(\mathbf{q})C(\mathbf{q}) = 0$$

Hence there exists a nonzero $E \in^{m\times m} [\mathbf{q}]$ such that $E(\mathbf{q})C(\mathbf{q}) = 0_{m\times m}$, which contradicts the fact that $C(\mathbf{q})$ has full normal rank (see Fact H.1). □

Fact H.3 Let $u\{1,N\}$ have a degree of persistency of ℓ, where

$$u(k) = \begin{bmatrix} u_1(k) \\ \vdots \\ u_m(k) \end{bmatrix} \in \mathbb{R}^m \tag{H.3}$$

and $u_i\{1,N\}$ has a degree of persistency of ℓ_i. Furthermore, let

$$i \geq \max(\ell_1, \ldots, \ell_m) - \ell \tag{H.4}$$

and let

$$C(\mathbf{q})u(k) = 0_{m\times 1}, \qquad \text{for all } k = 1, \ldots, N-\ell \tag{H.5}$$

where $C \in \mathbb{R}^{m\times m}[\mathbf{q}]$ is nonzero, and

$$C(\mathbf{q}) \triangleq C_\ell \mathbf{q}^\ell + \cdots + C_1\mathbf{q} + C_0 \tag{H.6}$$

Then the following statements are equivalent:

(i) $u\{1, N-i-1\}$ has a degree of persistency of ℓ, and C_ℓ has full rank.

(ii) For every nonzero $D \in \mathbb{R}^{m\times m}[\mathbf{q}]$ of degree less than or equal to $\ell + i$ which satisfies

$$D(\mathbf{q})u(k) = 0_{m\times 1}, \qquad \text{for all } k = 1, \ldots, N-\ell-i \tag{H.7}$$

there exists a nonzero $E \in \mathbb{R}^{m\times m}[\mathbf{q}]$ such that $D(\mathbf{q}) = E(\mathbf{q})C(\mathbf{q})$.

Proof First, suppose that (i) holds. Then since $C_\ell \in \mathbb{R}^{m\times m}$ has full rank, the matrix $\mathcal{T}_i(C)$ given by (12.20) has full row rank. Hence from Fact 12.9, it follows that (i)⇒(ii).

Next, suppose that (ii) holds, and let $\alpha_i \in \mathbb{R}[\mathbf{q}]$ denote the nonzero operator polynomial of degree less than or equal to ℓ_i such that

$$\alpha_i(\mathbf{q})u_i(k) = 0_{m\times 1}, \qquad \text{for all } k = 1, \ldots, N-\ell_i$$

Then letting $A \in \mathbb{R}^{m\times m}[\mathbf{q}]$ denote the diagonal operator polynomial

$$A(\mathbf{q}) \triangleq \begin{bmatrix} \alpha_1(\mathbf{q}) & & \\ & \ddots & \\ & & \alpha_m(\mathbf{q}) \end{bmatrix}$$

we have that $A(\mathbf{q})$ is a nonzero operator polynomial of degree less than or equal to xxxxxx which satisfies

$$A(\mathbf{q})u(k) = 0_{m\times 1}, \qquad \text{for all } k = 1, \ldots, N - \ell - i$$

Hence there exists a nonzero $E \in \mathbb{R}^{m\times m}[\mathbf{q}]$ such that $A(\mathbf{q}) = E(\mathbf{q})C(\mathbf{q})$. Specifically, since the leading coefficient of $A(\mathbf{q})$ has full rank ❒

References

[1] C. T. Chen, *Linear System Theory and Design*, 3rd ed. New York: Oxford University Press, 1999.

[2] E. Hairer, S. Norsett, and G. Wanner, *Solving Ordinary Differential Equations I: Nonstiff Problems*, 2nd ed. Springer-Verlag, 2000.

[3] C. F. Chan Man Fong, D. D. Kee, and P. N. Kaloni, *Advanced Mathematics for Engineering and Science.* World Scientific, 2003.

[4] W. A. Wolovich, *Linear Multivariable Systems.* New York, NY: Springer-Verlag, 1974.

[5] P. J. Antsaklis and A. N. Michel, *Linear Systems.* Boston: Birkhäuser, 2006.

[6] K. J. Åstrom and B. Wittenmark, *Computer-Controlled Systems.* Prentice-Hall, 1997.

[7] M. Verhaegen and V. Verdult, *Filtering and System Identification: A Least Squares Approach*, 1st ed. New York: Cambridge University Press, 2007.

[8] D. S. Bernstein, *Matrix Mathematics*, 2nd ed. Princeton, NJ: Princeton University Press, 2009.

[9] K. Ogata, *Modern Control Engineering*, 4th ed. Upper Saddle River, NJ: Prentice-Hall, 2002.

[10] J. N. Juang, *Applied System Identification.* Englewood Cliffs, NJ: Prentice-Hall, 1994.

[11] J. N. Juang and R. S. Pappa, “An eigensystem realization algorithm for modal parameter identification and model reduction,” *AIAA Journal of Guidance, Control, and Dynamics*, vol. 8, no. 5, pp. 620–627, 1985.

[12] L. Wasserman, *All of Nonparametric Statistics.* Springer-Verlag, 2006.

[13] A. Gut, *Probability: A Graduate Course.* Springer-Verlag, 2005.

[14] P. Billingsley, *Probability and Measure*, 3rd ed. Wiley, 1995.

[15] R. M. Gray, *Probability, Random Processes, and Ergodic Properties*, 2nd ed. New York: Springer-Verlag, 2009.

[16] P. J. Brockwell and R. A. Davis, *Time Series: Theory and Methods*, 2nd ed. New York: Springer-Verlag, 2006.

[17] L. Ljung, *System Identification: Theory for the User*, 3rd ed. Upper Saddle River, NJ: Prentice-Hall, 1999.

[18] R. Pintelon and J. Schoukens, *System Identification: A Frequency Domain Approach*, 1st ed. New York: Wiley-IEEE Press, 2001.

[19] J. N. Juang, M. Phan, L. G. Horta, and R. W. Longman, "Identification of observer/kalman filter markov parameters: Theory and experiments," *AIAA Journal of Guidance, Control, and Dynamics*, vol. 16, no. 2, pp. 320–329, April 1993.