CHAPTER 1
Introduction

1.1 WHY STOCHASTIC MODELS, ESTIMATION, AND CONTROL?

When considering system analysis or controller design, the engineer has at his disposal a wealth of knowledge derived from *deterministic* system and control theories. One would then naturally ask, why do we have to go beyond these results and propose *stochastic* system models, with ensuing concepts of estimation and control based upon these stochastic models? To answer this question, let us examine what the deterministic theories provide and determine where the shortcomings might be.

Given a physical system, whether it be an aircraft, a chemical process, or the national economy, an engineer first attempts to develop a mathematical model that adequately represents some aspects of the behavior of that system. Through physical insights, fundamental "laws," and empirical testing, he tries to establish the interrelationships among certain variables of interest, inputs to the system, and outputs from the system.

With such a mathematical model and the tools provided by system and control theories, he is able to investigate the system structure and modes of response. If desired, he can design compensators that alter these characteristics and controllers that provide appropriate inputs to generate desired system responses.

In order to observe the actual system behavior, measurement devices are constructed to output data signals proportional to certain variables of interest. These output signals and the known inputs to the system are the only information that is directly discernible about the system behavior. Moreover, if a feedback controller is being designed, the measurement device outputs are the only signals directly available for inputs to the controller.

There are three basic reasons why deterministic system and control theories do not provide a totally sufficient means of performing this analysis and

design. First of all, *no mathematical system model is perfect.* Any such model depicts only those characteristics of direct interest to the engineer's purpose. For instance, although an endless number of bending modes would be required to depict vehicle bending precisely, only a finite number of modes would be included in a useful model. The objective of the model is to represent the dominant or critical modes of system response, so many effects are knowingly left unmodeled. In fact, models used for generating online data processors or controllers must be pared to only the basic essentials in order to generate a computationally feasible algorithm.

Even effects which are modeled are necessarily *approximated* by a mathematical model. The "laws" of Newtonian physics are adequate approximations to what is actually observed, partially due to our being unaccustomed to speeds near that of light. It is often the case that such "laws" provide adequate system *structures*, but various *parameters* within that structure are not determined absolutely. Thus, there are many sources of uncertainity in any mathematical model of a system.

A second shortcoming of deterministic models is that dynamic systems are driven not only by our own control inputs, but also by *disturbances which we can neither control nor model deterministically.* If a pilot tries to command a certain angular orientation of his aircraft, the actual response will differ from his expectation due to wind buffeting, imprecision of control surface actuator responses, and even his inability to generate exactly the desired response from his own arms and hands on the control stick.

A final shortcoming is that *sensors do not provide perfect and complete data* about a system. First, they generally do not provide all the information we would like to know: either a device cannot be devised to generate a measurement of a desired variable or the cost (volume, weight, monetary, etc.) of including such a measurement is prohibitive. In other situations, a number of different devices yield functionally related signals, and one must then ask how to generate a best estimate of the variables of interest based on partially redundant data. Sensors do not provide exact readings of desired quantities, but introduce their own system dynamics and distortions as well. Furthermore, these devices are also always noise corrupted.

As can be seen from the preceding discussion, to assume perfect knowledge of all quantities necessary to describe a system completely and/or to assume perfect control over the system is a naive, and often inadequate, approach. This motivates us to ask the following four questions:

(1) How do you develop system models that account for these uncertainties in a direct and proper, yet practical, fashion?

(2) Equipped with such models and incomplete, noise-corrupted data from available sensors, how do you optimally estimate the quantities of interest to you?

(3) In the face of uncertain system descriptions, incomplete and noise-corrupted data, and disturbances beyond your control, how do you optimally control a system to perform in a desirable manner?

(4) How do you evaluate the performance capabilities of such estimation and control systems, both before and after they are actually built?

This book has been organized specifically to answer these questions in a meaningful and useful manner.

1.2 OVERVIEW OF THE TEXT

Chapters 2–4 are devoted to the stochastic modeling problem. First Chapter 2 reviews the pertinent aspects of deterministic system models, to be exploited and generalized subsequently. Probability theory provides the basis of all of our stochastic models, and Chapter 3 develops both the general concepts and the natural result of static system models. In order to incorporate dynamics into the model, Chapter 4 investigates stochastic processes, concluding with practical linear dynamic system models. The basic form is a linear system driven by white Gaussian noise, from which are available linear measurements which are similarly corrupted by white Gaussian noise. This structure is justified extensively, and means of describing a large class of problems in this context are delineated.

Linear estimation is the subject of the remaining chapters. Optimal filtering for cases in which a linear system model adequately describes the problem dynamics is studied in Chapter 5. With this background, Chapter 6 describes the design and performance analysis of practical online Kalman filters. Square root filters have emerged as a means of solving some numerical precision difficulties encountered when optimal filters are implemented on restricted word-length online computers, and these are detailed in Chapter 7.

Volume 1 is a complete text in and of itself. Nevertheless, Volume 2 will extend the concepts of linear estimation to smoothing, compensation of model inadequacies, system identification, and adaptive filtering. Nonlinear stochastic system models and estimators based upon them will then be fully developed. Finally, the theory and practical design of stochastic controllers will be described.

1.3 THE KALMAN FILTER: AN INTRODUCTION TO CONCEPTS

Before we delve into the details of the text, it would be useful to see where we are going on a conceptual basis. Therefore, the rest of this chapter will provide an overview of the optimal linear estimator, the Kalman filter. This will be conducted at a very elementary level but will provide insights into the

underlying concepts. As we progress through this overview, contemplate the ideas being presented: try to conceive of graphic *images* to portray the concepts involved (such as time propagation of density functions), and to generate a *logical structure* for the component pieces that are brought together to solve the estimation problem. If this basic conceptual framework makes sense to you, then you will better understand the need for the details to be developed later in the text. Should the idea of where we are going ever become blurred by the development of detail, refer back to this overview to regain sight of the overall objectives.

First one must ask, what is a Kalman filter? A Kalman filter is simply an *optimal recursive data processing algorithm.* There are many ways of defining *optimal,* dependent upon the criteria chosen to evaluate performance. It will be shown that, under the assumptions to be made in the next section, the Kalman filter is optimal with respect to virtually any criterion that makes sense. One aspect of this optimality is that the Kalman filter incorporates all information that can be provided to it. It processes all available measurements, regardless of their precision, to estimate the current value of the variables of interest, with use of (1) knowledge of the system and measurement device dynamics, (2) the statistical description of the system noises, measurement errors, and uncertainty in the dynamics models, and (3) any available information about initial conditions of the variables of interest. For example, to determine the velocity of an aircraft, one could use a Doppler radar, or the velocity indications of an inertial navigation system, or the pitot and static pressure and relative wind information in the air data system. Rather than ignore any of these outputs, a Kalman filter could be built to combine all of this data and knowledge of the various systems' dynamics to generate an overall best estimate of velocity.

The word *recursive* in the previous description means that, unlike certain data processing concepts, the Kalman filter does not require all previous data to be kept in storage and reprocessed every time a new measurement is taken. This will be of vital importance to the practicality of filter implementation.

The "filter" is actually a *data processing algorithm.* Despite the typical connotation of a filter as a "black box" containing electrical networks, the fact is that in most practical applications, the "filter" is just a computer program in a central processor. As such, it inherently incorporates discrete-time measurement samples rather than continuous time inputs.

Figure 1.1 depicts a typical situation in which a Kalman filter could be used advantageously. A system of some sort is driven by some known controls, and measuring devices provide the value of certain pertinent quantities. Knowledge of these system inputs and outputs is all that is explicitly available from the physical system for estimation purposes.

The *need* for a filter now becomes apparent. Often the variables of interest, some finite number of quantities to describe the "state" of the system, cannot

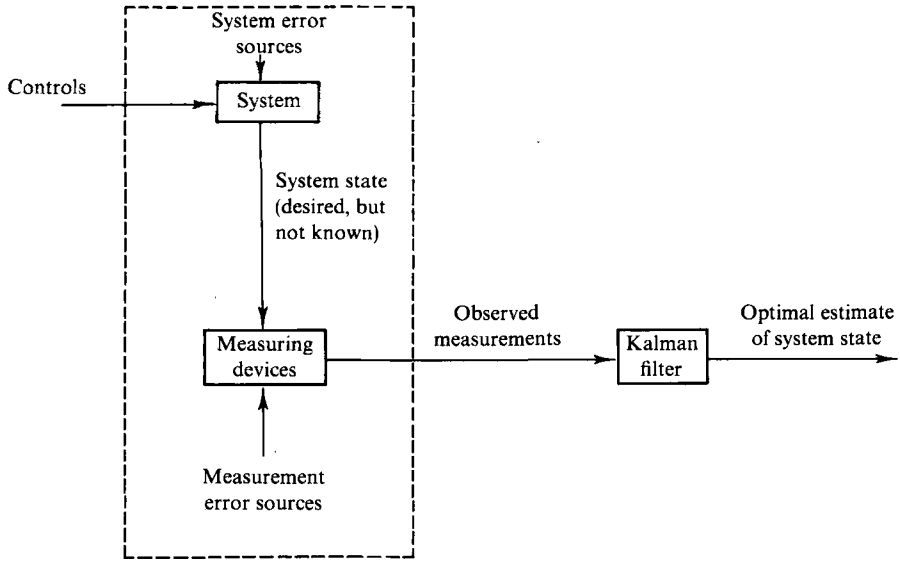

FIG. 1.1 Typical Kalman filter application.

be measured directly, and some means of inferring these values from the available data must be generated. For instance, an air data system directly provides static and pitot pressures, from which velocity must be inferred. This inference is complicated by the facts that the system is typically driven by inputs other than our own known controls and that the relationships among the various "state" variables and measured outputs are known only with some degree of uncertainty. Furthermore, any measurement will be corrupted to some degree by noise, biases, and device inaccuracies, and so a means of extracting valuable information from a noisy signal must be provided as well. There may also be a number of different measuring devices, each with its own particular dynamics and error characteristics, that provide some information about a particular variable, and it would be desirable to combine their outputs in a systematic and optimal manner. A Kalman filter combines all available measurement data, plus prior knowledge about the system and measuring devices, to produce an estimate of the desired variables in such a manner that the error is minimized *statistically*. In other words, if we were to run a number of candidate filters many times for the same application, then the average results of the Kalman filter would be better than the average results of any other.

Conceptually, what any type of filter tries to do is obtain an "optimal" estimate of desired quantities from data provided by a noisy environment, "optimal" meaning that it minimizes errors in some respect. There are many means of accomplishing this objective. If we adopt a Bayesian viewpoint, then we want the filter to propagate the *conditional probability density* of

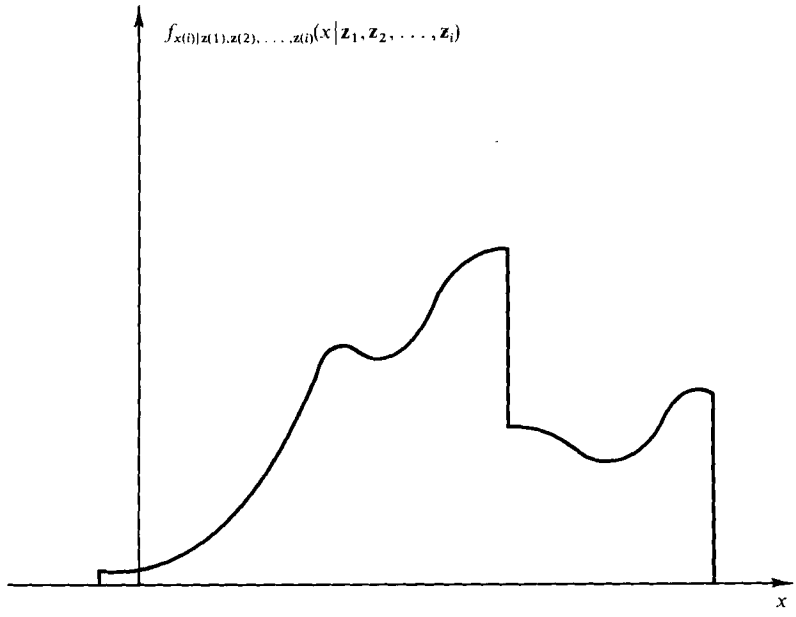

FIG. 1.2 Conditional probability density.

the desired quantities, conditioned on knowledge of the actual data coming from the measuring devices. To understand this concept, consider Fig. 1.2, a portrayal of a conditional probability density of the value of a scalar quantity x at time instant i ($x(i)$), conditioned on knowledge that the vector measurement $\mathbf{z}(1)$ at time instant 1 took on the value $\mathbf{z}_1$ ($\mathbf{z}(1) = \mathbf{z}_1$) and similarly for instants 2 through i, plotted as a function of possible $x(i)$ values. This is denoted as $f_{x(i)|\mathbf{z}(1),\mathbf{z}(2),\ldots,\mathbf{z}(i)}(x|\mathbf{z}_1,\mathbf{z}_2,\ldots,\mathbf{z}_i)$. For example, let $x(i)$ be the one-dimensional position of a vehicle at time instant i, and let $\mathbf{z}(j)$ be a two-dimensional vector describing the measurements of position at time j by two separate radars. Such a conditional probability density contains all the available information about $x(i)$: it indicates, for the given value of all measurements taken up through time instant i, what the probability would be of $x(i)$ assuming any particular value or range of values.

It is termed a "conditional" probability density because its shape and location on the x axis is dependent upon the values of the measurements taken. Its shape conveys the amount of certainty you have in the knowledge of the value of x. If the density plot is a narrow peak, then most of the probability "weight" is concentrated in a narrow band of x values. On the other hand, if the plot has a gradual shape, the probability "weight" is spread over a wider range of x, indicating that you are less sure of its value.

Once such a conditional probability density function is propagated, the "optimal" estimate can be defined. Possible choices would include

(1) the *mean*—the "center of probability mass" estimate;
(2) the *mode*—the value of x that has the highest probability, locating the peak of the density; and
(3) the *median*—the value of x such that half of the probability weight lies to the left and half to the right of it.

A Kalman filter performs this conditional probability density propagation for problems in which the system can be described through a *linear* model and in which system and measurement noises are *white* and *Gaussian* (to be explained shortly). Under these conditions, the mean, mode, median, and virtually any reasonable choice for an "optimal" estimate all coincide, so there is in fact a unique "best" estimate of the value of x. Under these three restrictions, the Kalman filter can be shown to be the best filter of any conceivable form. Some of the restrictions can be relaxed, yielding a qualified optimal filter. For instance, if the Gaussian assumption is removed, the Kalman filter can be shown to be the best (minimum error variance) filter out of the class of linear unbiased filters. However, these three assumptions can be justified for many potential applications, as seen in the following section.

1.4 BASIC ASSUMPTIONS

At this point it is useful to look at the three basic assumptions in the Kalman filter formulation. On first inspection, they may appear to be overly restrictive and unrealistic, To allay any misgivings of this sort, this section will briefly discuss the physical implications of these assumptions.

A linear system model is justifiable for a number of reasons. Often such a model is adequate for the purpose at hand, and when nonlinearities do exist, the typical engineering approach is to linearize about some nominal point or trajectory, achieving a perturbation model or error model. Linear systems are desirable in that they are more easily manipulated with engineering tools, and linear system (or differential equation) theory is much more complete and practical than nonlinear. The fact is that there are means of extending the Kalman filter concept to some nonlinear applications or developing nonlinear filters directly, but these are considered only if linear models prove inadequate.

"Whiteness" implies that the noise value is not correlated in time. Stated more simply, if you know what the value of the noise is now, this knowledge does you no good in predicting what its value will be at any other time. Whiteness also implies that the noise has equal power at all frequencies. Since this results in a noise with infinite power, a white noise obviously cannot really exist. One might then ask, why even consider such a concept if it does not

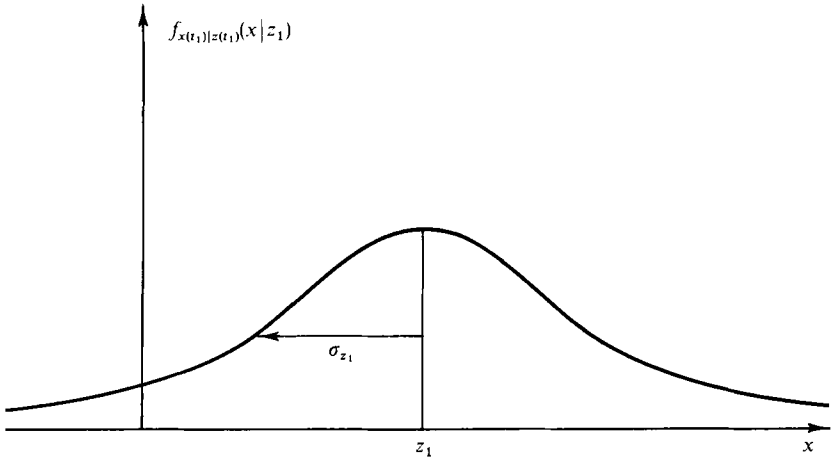

FIG. 1.4 Conditional density of position based on measured value z_1.

being in any one location, based upon the measurement you took. Note that σ_{z_1} is a direct measure of the uncertainty: the larger σ_{z_1} is, the broader the probability peak is, spreading the probability "weight" over a larger range of x values. For a Gaussian density, 68.3% of the probability "weight" is contained within the band σ units to each side of the mean, the shaded portion in Fig. 1.4.

Based on this conditional probability density, the best estimate of your position is

$$\hat{x}(t_1) = z_1 \tag{1-1}$$

and the variance of the error in the estimate is

$$\sigma_x{}^2(t_1) = \sigma_{z_1}^2 \tag{1-2}$$

Note that $\hat{x}$ is both the mode (peak) and the median (value with $\frac{1}{2}$ of the probability weight to each side), as well as the mean (center of mass).

Now say a trained navigator friend takes an independent fix right after you do, at time $t_2 \cong t_1$ (so that the true position has not changed at all), and obtains a measurement z_2 with a variance $\sigma_{z_2}^2$. Because he has a higher skill, assume the variance in his measurement to be somewhat smaller than in yours. Figure 1.5 presents the conditional density of your position at time t_2, based only on the measured value z_2. Note the narrower peak due to smaller variance, indicating that you are rather certain of your position based on his measurement.

At this point, you have two measurements available for estimating your position. The question is, how do you combine these data? It will be shown subsequently that, based on the assumptions made, the conditional density of

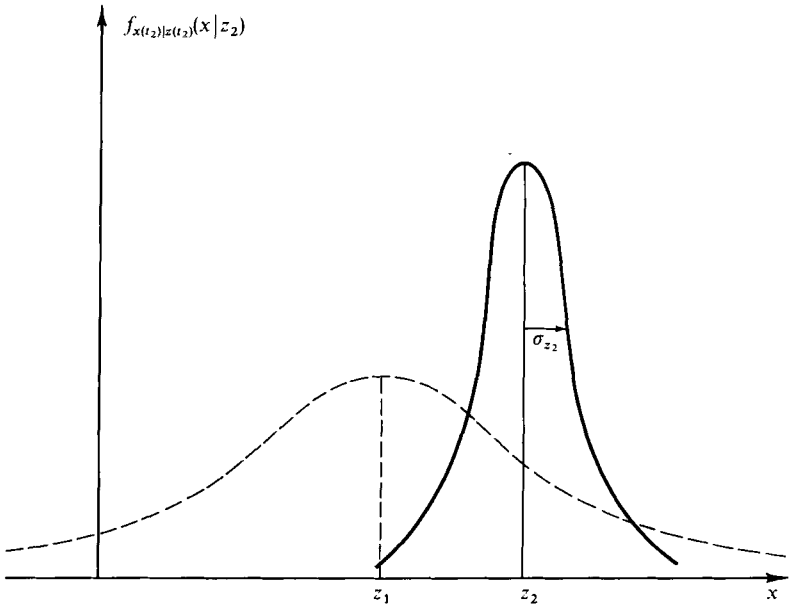

FIG. 1.5 Conditional density of position based on measurement z_2 alone.

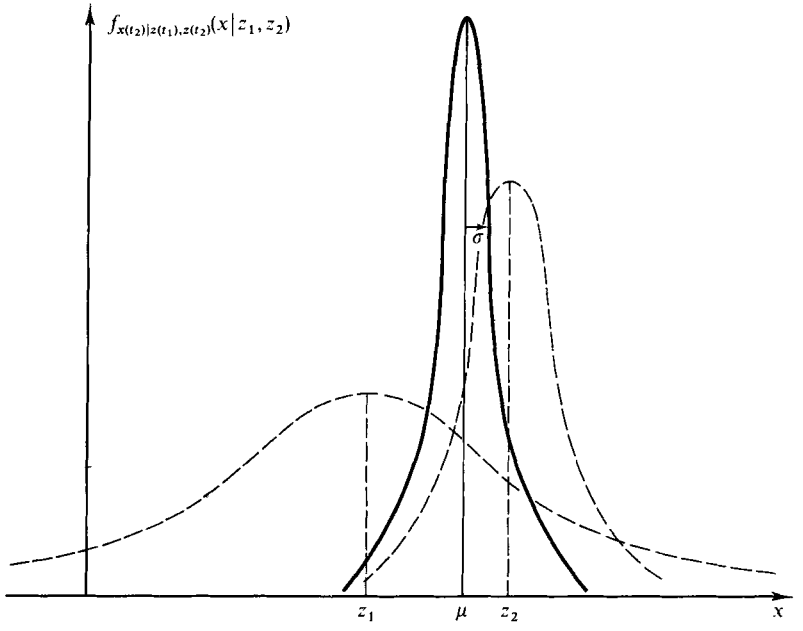

FIG. 1.6 Conditional density of position based on data z_1 and z_2.

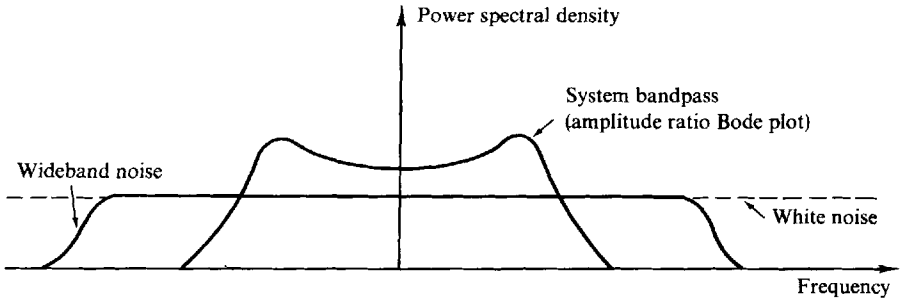

FIG. 1.3 Power spectral density bandwidths.

exist in real life? The answer is twofold. First, any physical system of interest has a certain frequency "bandpass"—a frequency range of inputs to which it can respond. Above this range, the input either has no effect, or the system so severely attentuates the effect that it essentially does not exist. In Fig. 1.3, a typical system bandpass curve is drawn on a plot of "power spectral density" (interpreted as the amount of power content at a certain frequency) versus frequency. Typically a system will be driven by wideband noise—one having power at frequencies above the system bandpass, and essentially constant power at all frequencies within the system bandpass—as shown in the figure. On this same plot, a white noise would merely extend this constant power level out across all frequencies. Now, within the bandpass of the system of interest, the fictitious white noise looks identical to the real wideband noise. So what has been gained? That is the second part of the answer to why a white noise model is used. It turns out that the mathematics involved in the filter are vastly simplified (in fact, made tractable) by replacing the real wideband noise with a white noise which, from the system's "point of view," is identical. Therefore, the white noise model is used.

One might argue that there are cases in which the noise power level is not constant over all frequencies within the system bandpass, or in which the noise is in fact time correlated. For such instances, a white noise put through a small linear system can duplicate virtually any form of time-correlated noise. This small system, called a "shaping filter," is then added to the original system, to achieve an overall linear system driven by white noise once again.

Whereas whiteness pertains to time or frequency relationships of a noise, Gaussianness has to do with its amplitude. Thus, at any single point in time, the probability density of a Gaussian noise amplitude takes on the shape of a normal bell-shaped curve. This assumption can be justified physically by the fact that a system or measurement noise is typically caused by a number of small sources. It can be shown mathematically that when a number of independent random variables are added together, the summed effect can be described very closely by a Gaussian probability density, regardless of the shape of the individual densities.

There is also a practical justification for using Gaussian densities. Similar to whiteness, it makes the mathematics tractable. But more than that, typically an engineer will know, at best, the first and second order statistics (mean and variance or standard deviation) of a noise process. In the absence of any higher order statistics, there is no better form to assume than the Gaussian density. The first and second order statistics completely determine a Gaussian density, unlike most densities which require an endless number of orders of statistics to specify their shape entirely. Thus, the Kalman filter, which propagates the first and second order statistics, includes *all* information contained in the conditional probability density, rather than only some of it, as would be the case with a different form of density.

The particular assumptions that are made are dictated by the objectives of, and the underlying motivation for, the model being developed. If our objective were merely to build good descriptive models, we would not confine our attention to linear system models driven by white Gaussian noise. Rather, we would seek the model, of whatever form, that best fits the data generated by the "real world." It is our desire to build estimators and controllers based upon our system models that drives us to these assumptions: other assumptions generally do not yield tractable estimation or control problem formulations. Fortunately, the class of models that yields tractable mathematics also provides adequate representations for many applications of interest. Later, the model structure will be extended somewhat to enlarge the range of applicability, but the requirement of model usefulness in subsequent estimator or controller design will again be a dominant influence on the manner in which the extensions are made.

1.5 A SIMPLE EXAMPLE

To see how a Kalman filter works, a simple example will now be developed. Any example of a single measuring device providing data on a single variable would suffice, but the determination of a position is chosen because the probability of one's exact location is a familiar concept that easily allows dynamics to be incorporated into the problem.

Suppose that you are lost at sea during the night and have no idea at all of your location. So you take a star sighting to establish your position (for the sake of simplicity, consider a one-dimensional location). At some time t_1 you determine your location to be z_1. However, because of inherent measuring device inaccuracies, human error, and the like, the result of your measurement is somewhat uncertain. Say you decide that the precision is such that the standard deviation (one-sigma value) involved is σ_{z_1} (or equivalently, the variance, or second order statistic, is $\sigma_{z_1}^2$). Thus, you can establish the conditional probability of $x(t_1)$, your position at time t_1, conditioned on the observed value of the measurement being z_1, as depicted in Fig. 1.4. This is a plot of $f_{x(t_1)|z(t_1)}(x|z_1)$ as a function of the location x: it tells you the probability of

your position at time $t_2 \cong t_1$, $x(t_2)$, given *both* z_1 *and* z_2, is a Gaussian density with mean μ and variance σ^2 as indicated in Fig. 1.6, with

$$\mu = [\sigma_{z_2}^2/(\sigma_{z_1}^2 + \sigma_{z_2}^2)]z_1 + [\sigma_{z_1}^2/(\sigma_{z_1}^2 + \sigma_{z_2}^2)]z_2 \tag{1-3}$$

$$1/\sigma^2 = (1/\sigma_{z_1}^2) + (1/\sigma_{z_2}^2) \tag{1-4}$$

Note that, from (1-4), σ is less than either σ_{z_1} or σ_{z_2}, which is to say that the uncertainty in your estimate of position has been decreased by combining the two pieces of information.

Given this density, the best estimate is

$$\hat{x}(t_2) = \mu \tag{1-5}$$

with an associated error variance σ^2. It is the mode and the mean (or, since it is the mean of a conditional density, it is also termed the conditional mean). Furthermore, it is also the maximum likelihood estimate, the weighted least squares estimate, and the linear estimate whose variance is less than that of any other linear unbiased estimate. In other words, it is the "best" you can do according to just about any reasonable criterion.

After some study, the form of μ given in Eq. (1-3) makes good sense. If σ_{z_1} were equal to σ_{z_2}, which is to say you think the measurements are of equal precision, the equation says the optimal estimate of position is simply the average of the two measurements, as would be expected. On the other hand, if σ_{z_1} were larger than σ_{z_2}, which is to say that the uncertainty involved in the measurement z_1 is greater than that of z_2, then the equation dictates "weighting" z_2 more heavily than z_1. Finally, the variance of the estimate is less than σ_{z_1} even if σ_{z_2} is very large: even poor quality data provide some information, and should thus increase the precision of the filter output.

The equation for $\hat{x}(t_2)$ can be rewritten as

$$\begin{aligned}\hat{x}(t_2) &= [\sigma_{z_2}^2/(\sigma_{z_1}^2 + \sigma_{z_2}^2)]z_1 + [\sigma_{z_1}^2/(\sigma_{z_1}^2 + \sigma_{z_2}^2)]z_2 \\ &= z_1 + [\sigma_{z_1}^2/(\sigma_{z_1}^2 + \sigma_{z_2}^2)][z_2 - z_1]\end{aligned} \tag{1-6}$$

or, in final form that is actually used in Kalman filter implementations [noting that $\hat{x}(t_1) = z_1$],

$$\hat{x}(t_2) = \hat{x}(t_1) + K(t_2)[z_2 - \hat{x}(t_1)] \tag{1-7}$$

where

$$K(t_2) = \sigma_{z_1}^2/(\sigma_{z_1}^2 + \sigma_{z_2}^2) \tag{1-8}$$

These equations say that the optimal estimate at time t_2, $\hat{x}(t_2)$, is equal to the best prediction of its value before z_2 is taken, $\hat{x}(t_1)$, plus a correction term of an optimal weighting value times the difference between z_2 and the best prediction of its value before it is actually taken, $\hat{x}(t_1)$. It is worthwhile to understand

this "predictor–corrector" structure of the filter. Based on all previous information, a prediction of the value that the desired variables and measurement will have at the next measurement time is made. Then, when the next measurement is taken, the difference between it and its predicted value is used to "correct" the prediction of the desired variables.

Using the $K(t_2)$ in Eq. (1-8), the variance equation given by Eq. (1-4) can be rewritten as

$$\sigma_x^2(t_2) = \sigma_x^2(t_1) - K(t_2)\sigma_x^2(t_1) \tag{1-9}$$

Note that the values of $\hat{x}(t_2)$ and $\sigma_x^2(t_2)$ embody all of the information in $f_{x(t_2)|z(t_1),\,z(t_2)}(x\,|\,z_1, z_2)$. Stated differently, by propagating these two variables, the conditional density of your position at time t_2, given z_1 and z_2, is completely specified.

Thus we have solved the static estimation problem. Now consider incorporating dynamics into the problem.

Suppose that you travel for some time before taking another measurement. Further assume that the best model you have of your motion is of the simple form

$$dx/dt = u + w \tag{1-10}$$

where u is a nominal velocity and w is a noise term used to represent the uncertainty in your knowledge of the actual velocity due to disturbances, off-nominal conditions, effects not accounted for in the simple first order equation, and the like. The "noise" w will be modeled as a white Gaussian noise with a mean of zero and variance of σ_w^2.

Figure 1.7 shows graphically what happens to the conditional density of position, given z_1 and z_2. At time t_2 it is as previously derived. As time progresses, the density travels along the x axis at the nominal speed u, while simultaneously spreading out about its mean. Thus, the probability density starts at the best estimate, moves according to the nominal model of dynamics,

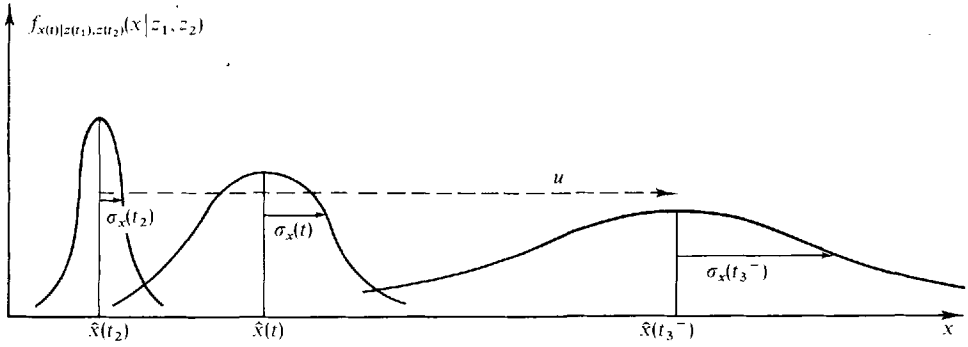

FIG. 1.7 Propagation of conditional probability density.

and spreads out in time because you become less sure of your exact position due to the constant addition of uncertainty over time. At the time t_3^-, just before the measurement is taken at time t_3, the density $f_{x(t_3)|z(t_1),z(t_2)}(x|z_1, z_2)$ is as shown in Fig. 1.7, and can be expressed mathematically as a Gaussian density with mean and variance given by

$$\hat{x}(t_3^-) = \hat{x}(t_2) + u[t_3 - t_2] \tag{1-11}$$

$$\sigma_x^2(t_3^-) = \sigma_x^2(t_2) + \sigma_w^2[t_3 - t_2] \tag{1-12}$$

Thus, $\hat{x}(t_3^-)$ is the optimal prediction of what the x value is at t_3^-, before the measurement is taken at t_3, and $\sigma_x^2(t_3^-)$ is the expected variance in that prediction.

Now a measurement is taken, and its value turns out to be z_3, and its variance is assumed to be $\sigma_{z_3}^2$. As before, there are now two Gaussian densities available that contain information about position, one encompassing all the information available before the measurement, and the other being the information provided by the measurement itself. By the same process as before, the density with mean $\hat{x}(t_3^-)$ and variance $\sigma_x^2(t_3^-)$ is combined with the density with mean z_3 and variance $\sigma_{z_3}^2$, to yield a Gaussian density with mean

$$\hat{x}(t_3) = \hat{x}(t_3^-) + K(t_3)[z_3 - \hat{x}(t_3^-)] \tag{1-13}$$

and variance

$$\sigma_x^2(t_3) = \sigma_x^2(t_3^-) - K(t_3)\sigma_x^2(t_3^-) \tag{1-14}$$

where the gain $K(t_3)$ is given by

$$K(t_3) = \sigma_x^2(t_3^-)/[\sigma_x^2(t_3^-) + \sigma_{z_3}^2] \tag{1-15}$$

The optimal estimate, $\hat{x}(t_3)$, satisfies the same form of equation as seen previously in (1-7). The best prediction of its value before z_3 is taken is corrected by an optimal weighting value times the difference between z_3 and the prediction of its value. Similarly, the variance and gain equations are of the same form as (1-8) and (1-9).

Observe the form of the equation for $K(t_3)$. If $\sigma_{z_3}^2$, the measurement noise variance, is large, then $K(t_3)$ is small; this simply says that you would tend to put little confidence in a very noisy measurement and so would weight it lightly. In the limit as $\sigma_{z_3}^2 \to \infty$, $K(t_3)$ becomes zero, and $\hat{x}(t_3)$ equals $\hat{x}(t_3^-)$: an infinitely noisy measurement is totally ignored. If the dynamic system noise variance σ_w^2 is large, then $\sigma_x^2(t_3^-)$ will be large [see Eq. (1-12)] and so will $K(t_3)$; in this case, you are not very certain of the output of the system model within the filter structure and therefore would weight the measurement heavily. Note that in the limit as $\sigma_w^2 \to \infty$, $\sigma_x^2(t_3^-) \to \infty$ and $K(t_3) \to 1$, so Eq. (1-13) yields

$$\hat{x}(t_3) = \hat{x}(t_3^-) + 1 \cdot [z_3 - \hat{x}(t_3^-)] = z_3 \tag{1-16}$$

Thus in the limit of absolutely no confidence in the system model output, the optimal policy is to ignore the output and use the new measurement as the optimal estimate. Finally, if $\sigma_x^2(t_3^-)$ should ever become zero, then so does $K(t_3)$; this is sensible since if $\sigma_x^2(t_3^-) = 0$, you are absolutely sure of your estimate before z_3 becomes available and therefore can disregard the measurement.

Although we have not as yet derived these results mathematically, we have been able to demonstrate the reasonableness of the filter structure.

1.6 A PREVIEW

Extending Eqs. (1-11) and (1-12) to the vector case and allowing time varying parameters in the system and noise descriptions yields the general Kalman filter algorithm for propagating the conditional density and optimal estimate from one measurement sample time to the next. Similarly, the Kalman filter update at a measurement time is just the extension of Eqs. (1-13)–(1-15). Further logical extensions would include estimation with data beyond the time when variables are to be estimated, estimation with nonlinear system models rather than linear, control of systems described through stochastic models, and both estimation and control when the noise and system parameters are not known with absolute certainty. The sequel provides a thorough investigation of those topics, developing both the theoretical mathematical aspects and practical engineering insights necessary to resolve the problem formulations and solutions fully.

GENERAL REFERENCES

The following references have influenced the development of both this introductory chapter and the entirety of this text.

1. Aoki, M., *Optimization of Stochastic Systems—Topics in Discrete-Time Systems*. Academic Press, New York, 1967.
2. Åström, K. J., *Introduction to Stochastic Control Theory*. Academic Press, New York, 1970.
3. Bryson, A. E. Jr., and Ho, Y., *Applied Optimal Control*. Blaisdell, Waltham, Massachusetts, 1969.
4. Bucy, R. S., and Joseph, P. D., *Filtering for Stochastic Processes with Applications to Guidance*. Wiley, New York, 1968.
5. Deutsch, R., *Estimation Theory*. Prentice-Hall, Englewood Cliffs, New Jersey, 1965.
6. Deyst, J. J., "Estimation and Control of Stochastic Processes," unpublished course notes. M.I.T. Dept. of Aeronautics and Astronautics, Cambridge, Massachusetts, 1970.
7. Gelb, A. (ed.), *Applied Optimal Estimation*. M.I.T. Press, Cambridge, Massachusetts, 1974.
8. Jazwinski, A. H., *Stochastic Processes and Filtering Theory*. Academic Press, New York, 1970.
9. Kwakernaak, H., and Sivan, R., *Linear Optimal Control Systems*. Wiley, New York, 1972.
10. Lee, R. C. K., *Optimal Estimation, Identification and Control*. M.I.T. Press, Cambridge, Massachusetts, 1964.
11. Liebelt, P. B., *An Introduction to Optimal Estimation*. Addison-Wesley, Reading, Massachusetts, 1967.
12. Maybeck, P. S., "The Kalman Filter—An Introduction for Potential Users," TM-72-3. Air Force Flight Dynamics Laboratory, Wright-Patterson AFB, Ohio, June 1972.

13. Maybeck, P. S., "Applied Optimal Estimation- -Kalman Filter Design and Implementation," notes for a continuing education course offered by the Air Force Institute of Technology, Wright-Patterson AFB, Ohio, semiannually since December 1974.
14. Meditch, J. S., *Stochastic Optimal Linear Estimation and Control*. McGraw-Hill, New York, 1969.
15. McGarty, T. P., *Stochastic Systems and State Estimation*. Wiley, New York, 1974.
16. Sage, A. P., and Melsa, J. L., *Estimation Theory with Application to Communications and Control*. McGraw-Hill, New York, 1971.
17. Schweppe, F. C., *Uncertain Dynamic Systems*. Prentice-Hall, Englewood Cliffs, New Jersey, 1973.
18. Van Trees, H. L., *Detection, Estimation and Modulation Theory*, Vol. 1. Wiley, New York, 1968.

APPENDIX AND PROBLEMS

Matrix Analysis

This appendix and its associated problems present certain results from elementary matrix analysis, as well as notation conventions, that will be of use throughout the text. If the reader desires more than this brief review, the list of references [1–11] at the end provides a partial list of good sources.

A.1 Matrices

An *n*-by-*m matrix* is a rectangular array of scalars consisting of *n* rows and *m* columns, denoted by a boldfaced capitalized letter, as

$$\mathbf{A} = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1m} \\ A_{21} & A_{22} & \cdots & A_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nm} \end{bmatrix}$$

Thus, A_{ij} is the scalar element in the ith row and jth column of $\mathbf{A}$, and unless specified otherwise, will be assumed herein to be a real number (or a real-valued scalar function).

If all of the elements A_{ij} are zeros, $\mathbf{A}$ is called a *zero matrix* or *null matrix*, denoted as $\mathbf{0}$.

If all of the elements of an *n*-by-*n* (*square*) matrix are zeros except for those along the principal diagonal, as

$$\mathbf{A} = \begin{bmatrix} A_{11} & 0 & \cdots & 0 \\ 0 & A_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_{nn} \end{bmatrix}$$

the $\mathbf{A}$ is called *diagonal*. Furthermore, if $A_{ii} = 1$ for all i, the matrix is called the *identity matrix* and is denoted by $\mathbf{I}$.

A square matrix is *symmetric* if $A_{ij} = A_{ji}$ for all values of i and j from 1 to n. Thus, a diagonal matrix is always symmetric. Show that there are at most $\frac{1}{2}n(n+1)$ nonredundant elements in an n-by-n symmetric matrix.

A *lower triangular* matrix is a square matrix, all of whose elements above the principal diagonal are zero, as

$$\mathbf{A} = \begin{bmatrix} A_{11} & 0 & \cdots & 0 \\ A_{21} & A_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nn} \end{bmatrix}$$

Similarly, an *upper triangular* matrix is a square matrix with all zeros below the principal diagonal.

A matrix composed of a single column, i.e., an n-by-1 matrix, is called an n-dimensional *vector* or *n-vector* and will be denoted by a boldfaced lower case letter, as

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Thus, x_i is the ith scalar element, or "component," of the n-vector $\mathbf{x}$. (The directed line segment from the origin to a point in Euclidean n-dimensional space can be represented, relative to a chosen basis or reference coordinate directions, by $\mathbf{x}$, and then x_i is the component along the ith basis vector or reference direction.) Properties of general nonsquare matrices (as described in Sections A.2, A.3, and A.10 to follow) are true specifically for vectors.

A matrix can be subdivided not only into its scalar elements, but also into arrays of elements called matrix *partitions*, such as

$$\mathbf{A} = \left[\begin{array}{c|c|c|c} \mathbf{A}_{11} & \mathbf{A}_{12} & \cdots & \mathbf{A}_{1i} \\ \vdots & \vdots & & \vdots \\ \hline \mathbf{A}_{k1} & \mathbf{A}_{k2} & \cdots & \mathbf{A}_{ki} \end{array}\right] \begin{matrix} \} n_1 \text{ rows} \\ \\ \} n_k \text{ rows} \end{matrix}$$

$$\begin{matrix} \underbrace{}_{m_1 \text{ columns}} & \underbrace{}_{m_2 \text{ columns}} & & \underbrace{}_{m_i \text{ columns}} \end{matrix}$$

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \hline \mathbf{x}_2 \\ \hline \vdots \\ \hline \mathbf{x}_k \end{bmatrix} \begin{matrix} \} n_1 \text{ components} \\ \} n_2 \text{ components} \\ \\ \} n_k \text{ components} \end{matrix}$$

A square matrix $\mathbf{A}$ is termed *block diagonal* if it can be subdivided into partitions such that $\mathbf{A}_{ij} = \mathbf{0}$ for all partitions for which $i \neq j$, and such that all partitions $\mathbf{A}_{ii}$ are square.

A.2 Equality, Addition, and Multiplication

Two n-by-m matrices $\mathbf{A}$ and $\mathbf{B}$ are *equal* if and only if $A_{ij} = B_{ij}$ for all i and j.

If $\mathbf{A}$ and $\mathbf{B}$ are both n-by-m matrices, their *sum* can be defined as $\mathbf{C} = \mathbf{A} + \mathbf{B}$, where $\mathbf{C}$ is an n-by-m matrix whose elements satisfy $C_{ij} = A_{ij} + B_{ij}$ for all i and j.

Their *difference* would be defined similarly. Show that

(a) $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$.
(b) $\mathbf{A} + (\mathbf{B} + \mathbf{C}) = (\mathbf{A} + \mathbf{B}) + \mathbf{C}$.
(c) $\mathbf{A} + \mathbf{0} = \mathbf{0} + \mathbf{A} = \mathbf{A}$.

The product of an n-by-m matrix $\mathbf{A}$ by a scalar b is the n-by-m matrix $\mathbf{C} = b\mathbf{A} = \mathbf{A}b$ composed of elements $C_{ij} = bA_{ij}$.

If $\mathbf{A}$ is n-by-m and $\mathbf{B}$ is m-by-r, then the *product* $\mathbf{C} = \mathbf{AB}$ can be defined as an n-by-r matrix with elements $C_{ij} = \sum_{k=1}^{m} A_{ik}B_{kj}$. This product can be defined only if the number of columns of $\mathbf{A}$ equals the number of rows of $\mathbf{B}$: only if $\mathbf{A}$ and $\mathbf{B}$ are "conformable" for the product $\mathbf{AB}$. Thus, the ordering in the product is important, and $\mathbf{AB}$ can be described as "premultiplying" $\mathbf{B}$ by $\mathbf{A}$ or "post-multiplying" $\mathbf{A}$ by $\mathbf{B}$. Show that for general conformable matrices

(d) $\mathbf{A(BC)} = \mathbf{(AB)C}$.
(e) $\mathbf{IA} = \mathbf{AI} = \mathbf{A}$.
(f) $\mathbf{0A} = \mathbf{A0} = \mathbf{0}$.
(g) $\mathbf{A(B + C)} = \mathbf{AB} + \mathbf{AC}$.
(h) in general, $\mathbf{AB} \neq \mathbf{BA}$, even for $\mathbf{A}$ and $\mathbf{B}$ both square.
(i) $\mathbf{AB} = \mathbf{0}$ in general does not imply that $\mathbf{A}$ or $\mathbf{B}$ is $\mathbf{0}$.

A product of particular importance is that of an n-by-m matrix $\mathbf{A}$ with an m-vector $\mathbf{x}$ to yield an n-vector $\mathbf{y} = \mathbf{Ax}$, with components $y_i = \sum_{j=1}^{m} A_{ij}x_j$. Such a matrix multiplication can be used to represent a linear transformation of a vector. More general functions, not expressible through matrix multiplications, can be written as

$$\mathbf{y} = \mathbf{f}(\mathbf{x}) \leftrightarrow \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} f_1(x_1, x_2, \ldots, x_m) \\ \vdots \\ f_n(x_1, x_2, \ldots, x_m) \end{bmatrix}$$

Matrix operations upon partitioned matrices obey the same rules of equality, addition, and multiplication, provided that the matrix partitions are conformable. For instance, show that

(j)

$$\left[\begin{array}{c|c} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \hline \mathbf{A}_{21} & \mathbf{A}_{22} \end{array}\right] + \left[\begin{array}{c|c} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \hline \mathbf{B}_{21} & \mathbf{B}_{22} \end{array}\right] = \left[\begin{array}{c|c} \mathbf{A}_{11} + \mathbf{B}_{11} & \mathbf{A}_{12} + \mathbf{B}_{12} \\ \hline \mathbf{A}_{21} + \mathbf{B}_{21} & \mathbf{A}_{22} + \mathbf{B}_{22} \end{array}\right]$$

(k)

$$\left[\begin{array}{c|c} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \hline \mathbf{A}_{21} & \mathbf{A}_{22} \end{array}\right] \cdot \left[\begin{array}{c|c} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \hline \mathbf{B}_{21} & \mathbf{B}_{22} \end{array}\right] = \left[\begin{array}{c|c} \mathbf{A}_{11}\mathbf{B}_{11} + \mathbf{A}_{12}\mathbf{B}_{21} & \mathbf{A}_{11}\mathbf{B}_{12} + \mathbf{A}_{12}\mathbf{B}_{22} \\ \hline \mathbf{A}_{21}\mathbf{B}_{11} + \mathbf{A}_{22}\mathbf{B}_{21} & \mathbf{A}_{21}\mathbf{B}_{12} + \mathbf{A}_{22}\mathbf{B}_{22} \end{array}\right]$$

A.3 Transposition

The *transpose* of an n-by-m matrix $\mathbf{A}$ is the m-by-n matrix denoted as $\mathbf{A}^T$ that satisfies $A_{ij}^T = A_{ji}$ for all i and j. Thus, transposition can be interpreted as interchanging the roles of rows and columns of a matrix. For example, if $\mathbf{x}$ is an n-vector, $\mathbf{x}^T$ is a 1-by-n matrix, or "row vector." Show that

(a) $(\mathbf{A}^T)^T = \mathbf{A}$.
(b) $(\mathbf{A} + \mathbf{B})^T = \mathbf{A}^T + \mathbf{B}^T$.
(c) $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$.
(d) if $\mathbf{A}$ is a symmetric matrix, $\mathbf{A}^T = \mathbf{A}$.
(e) if $\mathbf{x}$ and $\mathbf{y}$ are n-vectors, $\mathbf{x}^T\mathbf{y}$ is a scalar and $\mathbf{xy}^T$ is a square n-by-n matrix; $\mathbf{xx}^T$ is symmetric as well.
(f) if $\mathbf{A}$ is a symmetric n-by-n matrix and $\mathbf{B}$ is a general m-by-n matrix, then $\mathbf{C} = \mathbf{BAB}^T$ is a symmetric m-by-m matrix.
(g) if $\mathbf{A}$ and $\mathbf{B}$ are both symmetric n-by-n matrices, $(\mathbf{A} + \mathbf{B})$ is also symmetric but $(\mathbf{AB})$ generally is not.
(h)

$$\left[\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right]^T = \left[\begin{array}{c|c} \mathbf{A}^T & \mathbf{C}^T \\ \hline \mathbf{B}^T & \mathbf{D}^T \end{array}\right]$$

A.4 Matrix Inversion, Singularity, and Determinants

Given a square matrix $\mathbf{A}$, if there exists a matrix such that both premultiplying and postmultiplying it by $\mathbf{A}$ yields the identity, then this matrix is called the *inverse* of $\mathbf{A}$, and is denoted by $\mathbf{A}^{-1}$: $\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$. A square matrix that does not possess such an inverse is said to be *singular*. If $\mathbf{A}$ has an inverse, the inverse is unique, and $\mathbf{A}$ is termed *nonsingular*. Show that

(a) if $\mathbf{A}$ is nonsingular, then so is $\mathbf{A}^{-1}$, and $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$.
(b) $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$ if all indicated inverses exist.
(c) $(\mathbf{A}^{-1})^T = (\mathbf{A}^T)^{-1}$.
(d) if a transformation of variables is represented by $\mathbf{x}^* = \mathbf{Ax}$ and if $\mathbf{A}^{-1}$ exists, then $\mathbf{x} = \mathbf{A}^{-1}\mathbf{x}^*$.

The *determinant* of a square n-by-n matrix $\mathbf{A}$ is a scalar-valued function of the matrix elements, denoted by $|\mathbf{A}|$, the evaluation of which can be performed recursively through $|\mathbf{A}| = \sum_{j=1}^{n} A_{ij}C_{ij}$ for any fixed $i = 1, 2, \ldots,$ or n; C_{ij} is the "cofactor" of A_{ij}, defined through $C_{ij} = (-1)^{i+j}M_{ij}$, and M_{ij} is the "minor" of A_{ij}, defined as the determinant of the $(n-1)$-by-$(n-1)$ matrix formed by deleting the ith row and jth column of the n-by-n $\mathbf{A}$. (Note that iterative application of these relationships ends with the evaluation of determinants of 1-by-1 matrices or scalars as the scalar values themselves.) Show that

(e) if $\mathbf{A}$ is 2-by-2, then $|\mathbf{A}| = A_{11}A_{22} - A_{12}A_{21}$.

(f) if $\mathbf{A}$ is 3-by-3, then

$$|\mathbf{A}| = A_{11}A_{22}A_{33} + A_{12}A_{23}A_{31} + A_{13}A_{32}A_{21} - A_{11}A_{32}A_{23} - A_{12}A_{21}A_{33} - A_{13}A_{22}A_{31}$$

(g) $|\mathbf{A}^T| = |\mathbf{A}|$.

(h) if all the elements of any row or column of $\mathbf{A}$ are zero, $|\mathbf{A}| = 0$.

(i) if any row (column) of $\mathbf{A}$ is a multiple of any other row (column), then $|\mathbf{A}| = 0$.

(j) if a scalar multiple of any row (column) is added to any other row (column) of $\mathbf{A}$, the value of the determinant is unchanged.

(k) if $\mathbf{A}$ and $\mathbf{B}$ are n-by-n, $|\mathbf{AB}| = |\mathbf{A}|\,|\mathbf{B}|$.

(l) if $\mathbf{A}$ is diagonal, then $|\mathbf{A}|$ equals the product of its diagonal elements: $|\mathbf{A}| = \prod_{i=1}^{n} A_{ii}$.

(m) if the n-by-n $\mathbf{A}$ is nonsingular, then $|\mathbf{A}| \neq 0$ and $\mathbf{A}^{-1}$ can be evaluated as $\mathbf{A}^{-1} = [\text{adj}\,\mathbf{A}]/|\mathbf{A}|$, where $[\text{adj}\,\mathbf{A}]$ is the adjoint of $\mathbf{A}$, defined as the n-by-n matrix whose ij element (i.e., in the ith row and jth column) is the cofactor C_{ji}.

(n) $|\mathbf{A}^{-1}| = 1/|\mathbf{A}|$ if $|\mathbf{A}| \neq 0$.

(o) $|\mathbf{A}| = 0$ if and only if $\mathbf{A}$ is singular.

(p)

$$\left|\left[\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{0} & \mathbf{C} \end{array}\right]\right| = |\mathbf{A}|\,|\mathbf{C}|$$

If $\mathbf{A}$ is such that its inverse equals its transpose, $\mathbf{A}^{-1} = \mathbf{A}^T$, then $\mathbf{A}$ is termed *orthogonal*. If $\mathbf{A}$ is orthogonal, $\mathbf{AA}^T = \mathbf{A}^T\mathbf{A} = \mathbf{I}$, and $|\mathbf{A}| = \pm 1$.

A.5 Linear Independence and Rank

A set of k n-vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ is said to be *linearly dependent* if there exists a set of k constants $c_1, c_2, \ldots, c_k$ (at least one of which is not zero) such that $\sum_{i=1}^{k} c_i\mathbf{x}_i = \mathbf{0}$. If no such set of constants exists, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ are said to be *linearly independent*.

The *rank* of an n-by-m matrix is the order of the largest square nonsingular matrix that can be formed by deleting rows and columns. Show that

(a) if the m-by-n $\mathbf{A}$ is partitioned into column vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$ and $\mathbf{x}$ is an n-vector, then $\mathbf{Ax} = \sum_{i=1}^{n} \mathbf{a}_i x_i$.

(b) the rank of $\mathbf{A}$ equals the number of linearly independent rows or columns of $\mathbf{A}$, whichever is smaller.

(c) if $\mathbf{A}$ is n-by-n, then it is of rank n ("full rank") if and only if it is nonsingular.

(d) the rank of $\mathbf{xx}^T$ is one.

A.6 Eigenvalues and Eigenvectors

The equation $\mathbf{Ax} = \lambda\mathbf{x}$ for n-by-n $\mathbf{A}$, or $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$, possesses a nontrivial solution if an only if $|\mathbf{A} - \lambda\mathbf{I}| = 0$. The nth order polynomial $f(\lambda) = |\mathbf{A} - \lambda\mathbf{I}|$ is called the *characteristic polynomial* of $\mathbf{A}$, and the equation $f(\lambda) = 0$ is called its *characteristic equation.* The n *eigenvalues* of $\mathbf{A}$ are the (not necessarily distinct) roots of this equation, and the nonzero solutions to $\mathbf{Ax}_i = \lambda_i\mathbf{x}_i$, corresponding to the roots λ_i, are called *eigenvectors.* It can be shown that $|\mathbf{A}|$ equals the product of the eigenvalues of $\mathbf{A}$, and $\sum_{i=1}^{n} A_{ii} = \sum_{i=1}^{n} \lambda_i$.

Let the eigenvalues of the n-by-n $\mathbf{A}$ be the distinct values $\lambda_1, \lambda_2, \ldots, \lambda_n$, and let the associated eigenvectors be $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$. Then, if $\mathbf{E} = [\mathbf{e}_1|\mathbf{e}_2|\cdots|\mathbf{e}_n]$, $\mathbf{E}$ is nonsingular, and $\mathbf{E}^{-1}\mathbf{AE}$ is a diagonal matrix whose ith diagonal element is λ_i, for $i = 1, 2, \ldots, n$. Moreover, if $\mathbf{A}$ is also symmetric, then the eigenvalues are all real and $\mathbf{E}$ is orthogonal.

(a) Obtain the eigenvalues and eigenvectors for a general 2-by-2 $\mathbf{A}$; generate $\mathbf{E}$ and $\mathbf{E}^{-1}\mathbf{AE}$.

(b) Repeat for a general symmetric 2-by-2 $\mathbf{A}$; show that λ_1 and λ_2 must be real, and that $\mathbf{E}$ is orthogonal.

(c) Show that $|\mathbf{A}| = 0$ if and only if at least one eigenvalue is zero.

A.7 Quadratic Forms and Positive (Semi-) Definiteness

If $\mathbf{A}$ is n-by-n and $\mathbf{x}$ is an n-vector, then the scalar quantity $\mathbf{x}^T\mathbf{Ax}$ is called a *quadratic form.* Show that

(a) $\mathbf{x}^T\mathbf{Ax} = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij}x_ix_j$.

(b) without loss of generality, $\mathbf{A}$ can always be considered to be symmetric, since if $\mathbf{A}$ is not symmetric, a symmetric matrix $\mathbf{B}$ can always be defined by

$$B_{ij} = \begin{cases} A_{ij} & i = j \\ \frac{1}{2}(A_{ij} + A_{ji}) & i \neq j \end{cases}$$

for i and j equal to 1, 2, ..., n, and then $\mathbf{x}^T\mathbf{Ax} = \mathbf{x}^T\mathbf{Bx}$.

(c) if $\mathbf{A}$ is diagonal, $\mathbf{x}^T\mathbf{Ax} = \sum_{i=1}^{n} A_{ii}x_i^2$.

If $\mathbf{x}^T\mathbf{Ax} > 0$ for all $\mathbf{x} \neq 0$, the quadratic form is said to be *positive definite*, as is the matrix $\mathbf{A}$ itself, often written notationally as $\mathbf{A} > \mathbf{0}$. If $\mathbf{x}^T\mathbf{Ax} \geq 0$ for all $\mathbf{x} \neq \mathbf{0}$, the quadratic form and matrix $\mathbf{A}$ are termed *positive semidefinite*, denoted as $\mathbf{A} \geq \mathbf{0}$. Furthermore, the notation $\mathbf{A} > \mathbf{B}$ ($\mathbf{A} \geq \mathbf{B}$) is meant to say that $(\mathbf{A} - \mathbf{B})$ is positive definite (semidefinite). Show that

(d) if $\mathbf{A}$ is positive definite, it is nonsingular, and its inverse $\mathbf{A}^{-1}$ is also positive definite.

(e) the symmetric **A** is positive definite if and only if its eigenvalues are all positive; **A** is positive semidefinite if and only if its eigenvalues are all positive or zero.

(f) if **A** is positive definite and **B** is positive semidefinite, (**A** + **B**) is positive definite.

A.8 Trace

The *trace* of an n-by-n matrix **A**, denoted as tr(**A**), is defined as the sum of the diagonal terms:

$$\text{tr}(\mathbf{A}) \triangleq \sum_{i=1}^{n} A_{ii}$$

Using this basic definition, show that

(a) $\text{tr}(\mathbf{A}) = \text{tr}(\mathbf{A}^T)$.

(b) $\text{tr}(\mathbf{A}_1 + \mathbf{A}_2) = \text{tr}(\mathbf{A}_1) + \text{tr}(\mathbf{A}_2)$.

(c) if **B** is n-by-m and **C** is m-by-n, so that **BC** is n-by-n and **CB** is m-by-m, then

$$\text{tr}(\mathbf{BC}) = \text{tr}(\mathbf{CB}) = \text{tr}(\mathbf{B}^T\mathbf{C}^T) = \text{tr}(\mathbf{C}^T\mathbf{B}^T)$$

(d) if **x** and **y** are n-vectors and **A** is n-by-n, then

$$\text{tr}(\mathbf{x}\mathbf{y}^T) = \text{tr}(\mathbf{x}^T\mathbf{y}) = \mathbf{x}^T\mathbf{y}$$
$$\text{tr}(\mathbf{A}\mathbf{x}\mathbf{y}^T) = \text{tr}(\mathbf{y}^T\mathbf{A}\mathbf{x}) = \mathbf{y}^T\mathbf{A}\mathbf{x} = \mathbf{x}^T\mathbf{A}^T\mathbf{y}$$

A.9 Similarity

If **A** and **B** are n-by-n and **T** is a nonsingular n-by-n matrix, and $\mathbf{A} = \mathbf{T}^{-1}\mathbf{B}\mathbf{T}$, then **A** and **B** are said to be related by a *similarity transformation*, or are simply termed *similar*. Show that

(a) if $\mathbf{A} = \mathbf{T}^{-1}\mathbf{B}\mathbf{T}$, then $\mathbf{B} = \mathbf{T}\mathbf{A}\mathbf{T}^{-1}$.

(b) if **A** and **B** are similar, their determinants, eigenvalues, eigenvectors, characteristic polynomials, and traces are equal; also if **A** is positive definite, so is **B** and vice versa.

A.10 Differentiation and Integration

Let **A** be an n-by-m matrix function of a scalar variable t, such as time. Then $d\mathbf{A}/dt \triangleq \dot{\mathbf{A}}(t)$ is defined as the n-by-m matrix with ij element as dA_{ij}/dt for all i and j; $\int \mathbf{A}(\tau)\, d\tau$ is defined similarly as a matrix composed of elements $\int A_{ij}(\tau)\, d\tau$. Derivatives and integrals of vectors are special cases of these definitions. Show

that

(a) $d[\mathbf{A}^{\mathrm{T}}(t)]/dt = [d\mathbf{A}(t)/dt]^{\mathrm{T}}$ and similarly for integration.
(b) $d[\mathbf{A}(t)\mathbf{B}(t)]/dt = \dot{\mathbf{A}}(t)\mathbf{B}(t) + \mathbf{A}(t)\dot{\mathbf{B}}(t)$.

Let the scalar s and the n-vector $\mathbf{x}$ be functions of the m-vector $\mathbf{v}$. By convention, the following derivative definitions are made:

$$\frac{\partial s}{\partial \mathbf{v}} = \begin{bmatrix} \dfrac{\partial s}{\partial v_1} & \dfrac{\partial s}{\partial v_2} & \cdots & \dfrac{\partial s}{\partial v_m} \end{bmatrix}$$

$$\frac{\partial \mathbf{x}}{\partial \mathbf{v}} = \begin{bmatrix} \dfrac{\partial \mathbf{x}}{\partial v_1} & \dfrac{\partial \mathbf{x}}{\partial v_2} & \cdots & \dfrac{\partial \mathbf{x}}{\partial v_m} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial x_1}{\partial v_1} & \dfrac{\partial x_1}{\partial v_2} & \cdots & \dfrac{\partial x_1}{\partial v_m} \\ \dfrac{\partial x_2}{\partial v_1} & \dfrac{\partial x_2}{\partial v_2} & \cdots & \dfrac{\partial x_2}{\partial v_m} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial x_n}{\partial v_1} & \dfrac{\partial x_n}{\partial v_2} & \cdots & \dfrac{\partial x_n}{\partial v_m} \end{bmatrix}$$

By generating the appropriate forms for scalar components and recombining, show the validity of the following useful forms (for the vectors $\mathbf{x}$ and $\mathbf{y}$ assumed to be functions of $\mathbf{v}$ possibly, and the vector $\mathbf{z}$ and matrices $\mathbf{A}$ and $\mathbf{B}$ assumed constant):

(c) $\partial \mathbf{v}/\partial \mathbf{v} = \mathbf{I}$.
(d) $\partial(\mathbf{A}\mathbf{x})/\partial \mathbf{v} = \mathbf{A}\,\partial \mathbf{x}/\partial \mathbf{v}$, and thus, $\partial(\mathbf{A}\mathbf{v})/\partial \mathbf{v} = \mathbf{A}$.
(e) $\partial(\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{y})/\partial \mathbf{v} = \mathbf{x}^{\mathrm{T}}\mathbf{A}\,\partial \mathbf{y}/\partial \mathbf{v} + \mathbf{y}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\,\partial \mathbf{x}/\partial \mathbf{v}$

and so

$$\partial(\mathbf{z}^{\mathrm{T}}\mathbf{A}\mathbf{v})/\partial \mathbf{v} = \mathbf{z}^{\mathrm{T}}\mathbf{A}, \qquad \partial(\mathbf{v}^{\mathrm{T}}\mathbf{A}\mathbf{z})/\partial \mathbf{v} = \mathbf{z}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}$$

and

$$\partial(\mathbf{v}^{\mathrm{T}}\mathbf{A}\mathbf{v})/\partial \mathbf{v} = \mathbf{v}^{\mathrm{T}}\mathbf{A} + \mathbf{v}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}} = 2\mathbf{v}^{\mathrm{T}}\mathbf{A} \qquad \text{if} \quad \mathbf{A} = \mathbf{A}^{\mathrm{T}}$$

and

$$\partial\{(\mathbf{z} - \mathbf{B}\mathbf{v})^{\mathrm{T}}\mathbf{A}(\mathbf{z} - \mathbf{B}\mathbf{v})\}/\partial \mathbf{v} = -2(\mathbf{z} - \mathbf{B}\mathbf{v})^{\mathrm{T}}\mathbf{A}\mathbf{B} \qquad \text{if} \quad \mathbf{A} = \mathbf{A}^{\mathrm{T}}.$$

REFERENCES

1. Bellman, R. E., *Introduction to Matrix Analysis*. McGraw-Hill, New York, 1960.
2. DeRusso, P. M., Roy, R. J., and Close, C. M., *State Variables for Engineers*. Wiley, New York, 1965.
3. Edelen, D. G. B., and Kydoniefs, A. D., *An Introduction to Linear Algebra for Science and Engineering* (2nd Ed.). American Elsevier, New York, 1976.

4. Gantmacher, R. F., *Matrix Theory*, Vol. I, Chelsea, New York, 1959 (translation of 1953 Russian edition).
5. Hildebrand, F. B., *Methods of Applied Mathematics* (2nd Ed.). Prentice-Hall, Englewood Cliffs, New Jersey, 1965.
6. Hoffman, K., and Kunze, R., *Linear Algebra*. Prentice-Hall, Englewood Cliffs, New Jersey, 1961.
7. Nering, E. D., *Linear Algebra and Matrix Theory* (2nd Ed.). Wiley, New York, 1970.
8. Noble, B., *Applied Linear Algebra*. Prentice-Hall, Englewood Cliffs, New Jersey, 1969.
9. Ogata, K., *State Space Analysis of Control Systems*. Prentice-Hall, Englewood Cliffs, New Jersey, 1967.
10. Polak, E., and Wong, E., *Notes for a First Course on Linear Systems*. Van Nostrand-Reinhold, Princeton, New Jersey, 1970.
11. Zadeh, L. A., and Desoer, C. A., *Linear System Theory*. McGraw-Hill, New York, 1963.