Part I

COMMUNICATION and REGENERATION

Chapter 1

Heuristics

This book is about Markovian models, and particularly about the structure and stability of such models. We develop a theoretical basis by studying Markov chains in very general contexts; and we develop, as systematically as we can, the applications of this theory to applied models in systems engineering, in operations research, and in time series.

A Markov chain is, for us, a collection of random variables $\mathbf{\Phi} = \{\Phi_n : n \in T\}$, where T is a countable time set. It is customary to write T as $\mathbb{Z}_+ := \{0, 1, \ldots\}$, and we will do this henceforth.

Heuristically, the critical aspect of a Markov model, as opposed to any other set of random variables, is that it is forgetful of all but its most immediate past. The precise meaning of this requirement for the evolution of a Markov model in time, that the future of the process is independent of the past given only its present value, and the construction of such a model in a rigorous way, is taken up in Chapter 3. Until then it is enough to indicate that for a process $\mathbf{\Phi}$, evolving on a space X and governed by an overall probability law P, to be a time-homogeneous Markov chain, there must be a set of "transition probabilities" $\{P^n(x, A), x \in \mathsf{X}, A \subset \mathsf{X}\}$ for appropriate sets A such that for times n, m in $\mathbb{Z}_+$

$$\mathsf{P}(\Phi_{n+m} \in A \mid \Phi_j, j \le m; \Phi_m = x) = P^n(x, A); \tag{1.1}$$

that is, $P^n(x, A)$ denotes the probability that a chain at x will be in the set A after n steps, or transitions. The independence of P^n on the values of $\Phi_j, j \le m$, is the Markov property, and the independence of P^n and m is the time-homogeneity property.

We now show that systems which are amenable to modeling by discrete time Markov chains with this structure occur frequently, especially if we take the state space of the process to be rather general, since then we can allow auxiliary information on the past to be incorporated to ensure the Markov property is appropriate.

1.1 A range of Markovian environments

The following examples illustrate this breadth of application of Markov models, and a little of the reason why stability is a central requirement for such models.

(a) The cruise control system on a modern motor vehicle monitors, at each time point k, a vector $\{X_k\}$ of inputs: speed, fuel flow, and the like (see Kuo [230]). It calculates a control value U_k which adjusts the throttle, causing a change in the values of the environmental variables X_{k+1} which in turn causes U_{k+1} to change again. The multidimensional process $\Phi_k = \{X_k, U_k\}$ is often a Markov chain (see Section 2.3.2), with new values overriding those of the past, and with the next value governed by the present value. All of this is subject to measurement error, and the process can never be other than stochastic: stability for this chain consists in ensuring that the environmental variables do not deviate too far, within the limits imposed by randomness, from the pre-set goals of the control algorithm.

(b) A queue at an airport evolves through the random arrival of customers and the service times they bring. The numbers in the queue, and the time the customer has to wait, are critical parameters for customer satisfaction, for waiting room design, for counter staffing (see Asmussen [9]). Under appropriate conditions (see Section 2.4.2), variables observed at arrival times (either the queue numbers, or a combination of such numbers and aspects of the remaining or currently uncompleted service times) can be represented as a Markov chain, and the question of stability is central to ensuring that the queue remains at a viable level. Techniques arising from the analysis of such models have led to the now familiar single-line multi-server counters actually used in airports, banks and similar facilities, rather than the previous multi-line systems.

(c) The exchange rate X_n between two currencies can be and is represented as a function of its past several values $X_{n-1}, \ldots, X_{n-k}$, modified by the volatility of the market which is incorporated as a disturbance term W_n (see Krugman and Miller [222] for models of such fluctuations). The autoregressive model

$$X_n = \sum_{j=1}^{k} \alpha_j X_{n-j} + W_n$$

central in time series analysis (see Section 2.1) captures the essential concept of such a system. By considering the whole k-length vector $\Phi_n = (X_n, \ldots, X_{n-k+1})$, Markovian methods can be brought to the analysis of such time-series models. Stability here involves relatively small fluctuations around a norm; and as we will see, if we do not have such stability, then typically we will have instability of the grossest kind, with the exchange rate heading to infinity.

(d) Storage models are fundamental in engineering, insurance and business. In engineering one considers a dam, with input of random amounts at random times, and a steady withdrawal of water for irrigation or power usage. This model has a Markovian representation (see Section 2.4.3 and Section 2.4.4). In insurance, there is a steady inflow of premiums, and random outputs of claims at random times. This model is also a storage process, but with the input and output reversed when compared to the engineering version, and also has a Markovian representation (see Asmussen [9]). In business, the inventory of a firm will act in a manner between these two models, with regular but sometimes also large irregular withdrawals,

and irregular ordering or replacements, usually triggered by levels of stock reaching threshold values (for an early but still relevant overview see Prabhu [322]). This also has, given appropriate assumptions, a Markovian representation. For all of these, stability is essentially the requirement that the chain stays in "reasonable values": the stock does not overfill the warehouse, the dam does not overflow, the claims do not swamp the premiums.

(e) The growth of populations is modeled by Markov chains, of many varieties. Small homogeneous populations are branching processes (see Athreya and Ney [12]); more coarse analysis of large populations by time series models allows, as in (c), a Markovian representation (see Brockwell and Davis [51]); even the detailed and intricate cycle of the Canadian lynx seem to fit a Markovian model [287], [388]. Of these, only the third is stable in the sense of this book: the others either die out (which is, trivially, stability but a rather uninteresting form); or, as with human populations, expand (at least within the model) forever.

(f) Markov chains are currently enjoying wide popularity through their use as a tool in simulation: Gibbs sampling, and its extension to Markov chain Monte Carlo methods of simulation, which utilise the fact that many distributions can be constructed as invariant or limiting distributions (in the sense of (1.16) below), has had great impact on a number of areas (see, as just one example, [312]). In particular, the calculation of posterior Bayesian distributions has been revolutionized through this route [359, 381, 385], and the behavior of prior and posterior distributions on very general spaces such as spaces of likelihood measures themselves can be approached in this way (see [112]): there is no doubt that at this degree of generality, techniques such as we develop in this book are critical.

(g) There are Markov models in all areas of human endeavor. The degree of word usage by famous authors admits a Markovian representation (see, amongst others, Gani and Saunders [136]). Did Shakespeare have an unlimited vocabulary? This can be phrased as a question of stability: if he wrote forever, would the size of the vocabulary used grow in an unlimited way? The record levels in sport are Markovian (see Resnick [325]). The spread of surnames may be modeled as Markovian (see [78]). The employment structure in a firm has a Markovian representation (see Bartholomew and Forbes [18]). This range of examples does not imply all human experience is Markovian: it does indicate that if enough variables are incorporated in the definition of "immediate past", a forgetfulness of all but that past is a reasonable approximation, and one which we can handle.

(h) Perhaps even more importantly, at the current level of technological development, telecommunications and computer networks have inherent Markovian representations (see Kelly [199] for a very wide range of applications, both actual and potential, and Gray [144] for applications to coding and information theory). They may be composed of sundry connected queueing processes, with jobs completed at nodes, and messages routed between them; to summarize the past one may need a state space which is the product of many subspaces, including countable subspaces, representing numbers in queues and buffers, uncountable subspaces, representing unfinished service times or routing times, or numerous trivial 0-1 subspaces representing available slots or wait-states or busy servers. But by a suitable choice of

state space, and (as always) a choice of appropriate assumptions, the methods we give in this book become tools to analyze the stability of the system.

Simple spaces do not describe these systems in general. Integer or real-valued models are sufficient only to analyze the simplest models in almost all of these contexts.

The methods and descriptions in this book are for chains which take their values in a virtually arbitrary space X. We do not restrict ourselves to countable spaces, nor even to Euclidean space $\mathbb{R}^n$, although we do give specific formulations of much of our theory in both these special cases, to aid both understanding and application.

One of the key factors that allows this generality is that, for the models we consider, there is no great loss of power in going from a simple to a quite general space. The reader interested in any of the areas of application above should therefore find that the structural and stability results for general Markov chains are potentially tools of great value, no matter what the situation, no matter how simple or complex the model considered.

1.2 Basic models in practice

1.2.1 The Markovian assumption

The simplest Markov models occur when the variables Φ_n, $n \in \mathbb{Z}_+$, are independent. However, a collection of random variables which is independent certainly fails to capture the essence of Markov models, which are designed to represent systems which *do* have a past, even though they depend on that past only through knowledge of the most recent information on their trajectory.

As we have seen in Section 1.1, the seemingly simple Markovian assumption allows a surprisingly wide variety of phenomena to be represented as Markov chains. It is this which accounts for the central place that Markov models hold in the stochastic process literature. For once some limited independence of the past is allowed, then there is the possibility of reformulating many models so the dependence is as simple as in (1.1).

There are two standard paradigms for allowing us to construct Markovian representations, even if the initial phenomenon appears to be non-Markovian.

In the first, the dependence of some model of interest $\boldsymbol{Y} = \{Y_n\}$ on its past values may be non-Markovian but still be based only on a finite "memory". This means that the system depends on the past only through the previous $k+1$ values, in the probabilistic sense that

$$\mathsf{P}(Y_{n+m} \in A \mid Y_j, j \le n) = \mathsf{P}(Y_{n+m} \in A \mid Y_j, j = n, n-1, \ldots, n-k). \tag{1.2}$$

Merely by reformulating the model through defining the vectors

$$\Phi_n = \{Y_n, \ldots, Y_{n-k}\}$$

and setting $\boldsymbol{\Phi} = \{\Phi_n, n \ge 0\}$ (taking obvious care in defining $\{\Phi_0, \ldots, \Phi_{k-1}\}$), we can define from $\boldsymbol{Y}$ a Markov chain $\boldsymbol{\Phi}$. The motion in the first coordinate of $\boldsymbol{\Phi}$ reflects that of $\boldsymbol{Y}$, and in the other coordinates is trivial to identify, since Y_n becomes $Y_{(n+1)-1}$, and so forth; and hence $\boldsymbol{Y}$ can be analyzed by Markov chain methods.

Such *state space* representations, despite their somewhat artificial nature in some cases, are an increasingly important tool in deterministic and stochastic systems theory, and in linear and nonlinear time series analysis.

As the second paradigm for constructing a Markov model representing a non-Markovian system, we look for so-called *embedded regeneration points*. These are times at which the system forgets its past in a probabilistic sense: the system viewed at such time points is Markovian even if the overall process is not.

Consider as one such model a storage system, or dam, which fills and empties. This is rarely Markovian: for instance, knowledge of the time since the last input, or the size of previous inputs still being drawn down, will give information on the current level of the dam or even the time to the next input. But at that very special sequence of times when the dam is empty and an input actually occurs, the process may well "forget the past", or "regenerate": appropriate conditions for this are that the times between inputs and the size of each input are independent. For then one cannot forecast the time to the next input when at an input time, and the current emptiness of the dam means that there is no information about past input levels available at such times. The dam content, viewed at these special times, can then be analyzed as a Markov chain.

"Regenerative models" for which such "embedded Markov chains" occur are common in operations research, and in particular in the analysis of queueing and network models.

State space models and regeneration time representations have become increasingly important in the literature of time series, signal processing, control theory, and operations research, and not least because of the possibility they provide for analysis through the tools of Markov chain theory. In the remainder of this opening chapter, we will introduce a number of these models in their simplest form, in order to provide a concrete basis for further development.

1.2.2 State space and deterministic control models

One theme throughout this book will be the analysis of stochastic models through consideration of the underlying deterministic motion of specific (non-random) realizations of the input driving the model.

Such an approach draws on both control theory, for the deterministic analysis; and Markov chain theory, for the translation to the stochastic analogue of the deterministic chain.

We introduce both of these ideas heuristically in this section.

Deterministic control models

In the theory of deterministic systems and control systems we find the simplest possible Markov chains: ones such that the next position of the chain is determined completely as a function of the previous position.

Consider the deterministic linear system on $\mathbb{R}^n$, whose "state trajectory" $\boldsymbol{x} = \{x_k, k \in \mathbb{Z}_+\}$ is defined inductively as

$$x_{k+1} = Fx_k \tag{1.3}$$

where F is an $n \times n$ matrix.

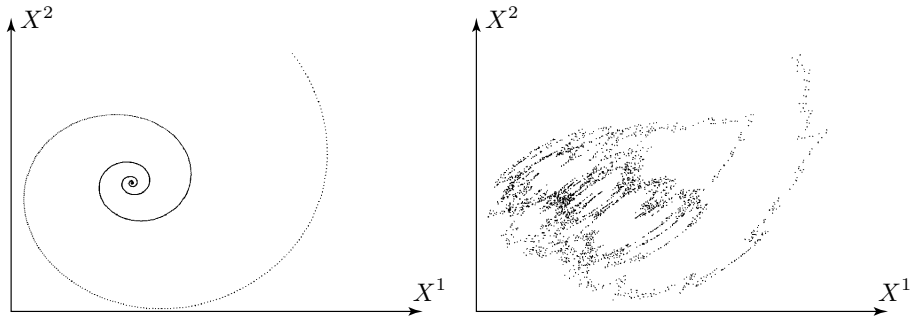

Figure 1.1: At left is a sample path generated by the deterministic linear model on $\mathbb{R}^2$. At right is a sample path from the linear state space model on $\mathbb{R}^2$ with Gaussian noise.

Clearly, this is a multidimensional Markovian model: even if we know all of the values of $\{x_k, k \le m\}$ then we will still predict x_{m+1} in the same way, with the same (exact) accuracy, based solely on (1.3) which uses only knowledge of x_m.

At left in Figure 1.1 we show a sample path corresponding to the choice of F as $F = I + \Delta A$ with I equal to a 2×2 identity matrix, $A = \left(\begin{smallmatrix} -0.2, & 1 \\ -1, & -0.2 \end{smallmatrix}\right)$ and $\Delta = 0.02$. It is instructive to realize that two very different types of behavior can follow from related choices of the matrix F. The trajectory spirals in, and is intuitively "stable"; but if we read the model in the other direction, the trajectory spirals out, and this is exactly the result of using F^{-1} in (1.3).

Thus, although this model is one without any built-in randomness or stochastic behavior, questions of stability of the model are still basic: the first choice of F gives a stable model, the second choice of F^{-1} gives an unstable model.

A straightforward generalization of the linear system of (1.3) is the *linear control model*. From the outward version of the trajectory in Figure 1.1, it is clearly possible for the process determined by F to be out of control in an intuitively obvious sense. In practice, one might observe the value of the process, and influence it either by adding on a modifying "control value" either independently of the current position of the process or directly based on the current value. Now the state trajectory $\boldsymbol{x} = \{x_k\}$ on $\mathbb{R}^n$ is defined inductively not only as a function of its past, but also of such a (deterministic) control sequence $\boldsymbol{u} = \{u_k\}$ taking values in, say, $\mathbb{R}^p$.

Formally, we can describe the linear control model by the postulates (LCM1) and (LCM2) below.

If the control value u_{k+1} depends at most on the sequence $x_j, j \le k$ through x_k, then it is clear that the LCM(F,G) model is itself Markovian.

However, the interest in the linear control model in our context comes from the fact that it is helpful in studying an associated Markov chain called the *linear state space model*. This is simply (1.4) with a certain random choice for the sequence $\{u_k\}$, with u_{k+1} independent of $x_j, j \le k$, and we describe this next.

Deterministic linear control model

Suppose $\boldsymbol{x} = \{x_k\}$ is a process on $\mathbb{R}^n$ and $\boldsymbol{u} = \{u_n\}$ is a process on $\mathbb{R}^p$, for which x_0 is arbitrary and for $k \geq 1$

(LCM1) there exists an $n \times n$ matrix F and an $n \times p$ matrix G such that for each $k \in \mathbb{Z}_+$,

$$x_{k+1} = Fx_k + Gu_{k+1}; \tag{1.4}$$

(LCM2) the sequence $\{u_k\}$ on $\mathbb{R}^p$ is chosen deterministically.

Then $\boldsymbol{x}$ is called the *linear control model driven by* F, G, or the LCM(F,G) model.

The linear state space model

In developing a stochastic version of a control system, an obvious generalization is to assume that the next position of the chain is determined as a function of the previous position, but in some way which still allows for uncertainty in its new position, such as by a random choice of the "control" at each step. Formally, we can describe such a model by

Linear state space model

Suppose $\boldsymbol{X} = \{X_k\}$ is a stochastic process for which

(LSS1) there exists an $n \times n$ matrix F and an $n \times p$ matrix G such that for each $k \in \mathbb{Z}_+$, the random variables X_k and W_k take values in $\mathbb{R}^n$ and $\mathbb{R}^p$, respectively, and satisfy inductively for $k \in \mathbb{Z}_+$,

$$X_{k+1} = FX_k + GW_{k+1}$$

where X_0 is arbitrary;

(LSS2) the random variables $\{W_k\}$ are independent and identically distributed (i.i.d), and are independent of X_0, with common distribution $\Gamma(A) = \mathsf{P}(W_j \in A)$ having finite mean and variance.

Then $\boldsymbol{X}$ is called the linear state space model driven by F, G, or the LSS(F,G) model, with *associated* control model LCM(F,G).

Such linear models with random "noise" or "innovation" are related to both the simple deterministic model (1.3) and also the linear control model (1.4).

There are obviously two components to the evolution of a state space model. The matrix F controls the motion in one way, but its action is modulated by the regular input of random fluctuations which involve both the underlying variable with distribution Γ, and its adjustment through G. At left in Figure 1.1 we show a sample path corresponding to the same matrix F, $G = \binom{2.5}{2.5}$, and with Γ taken as a bivariate Normal, or Gaussian, distribution $N(0,1)$. This indicates that the addition of the noise variables $\boldsymbol{W}$ can lead to types of behavior very different to that of the deterministic model, even with the same choice of the function F.

Such models describe the movements of airplanes, of industrial and engineering equipment, and even (somewhat idealistically) of economies and financial systems [3, 57]. Stability in these contexts is then understood in terms of return to level flight, or small and (in practical terms) insignificant deviations from set engineering standards, or minor inflation or exchange-rate variation. Because of the random nature of the noise we cannot expect totally unvarying systems; what we seek to preclude are explosive or wildly fluctuating operations.

We will see that, in wide generality, if the linear control model LCM(F,G) is stable in a deterministic way, and if we have a "reasonable" distribution Γ for our random control sequences, then the linear state space LSS(F,G) model is also stable in a stochastic sense.

In Chapter 2 we will describe models which build substantially on these simple structures, and which illustrate the development of Markovian structures for linear and nonlinear state space model theory.

We now leave state space models, and turn to the simplest examples of another class of models, which may be thought of collectively as models with a regenerative structure.

1.2.3 The gamblers ruin and the random walk

Unrestricted random walk

At the roots of traditional probability theory lies the problem of the gambler's ruin.

One has a gaming house in which one plays successive games; at each time point, there is a playing of a game, and an amount won or lost: and the successive totals of the amounts won or lost represent the fluctuations in the fortune of the gambler.

It is common, and realistic, to assume that as long as the gambler plays the same game each time, then the winnings W_k at each time k are i.i.d.

Now write the total winnings (or losings) at time k as Φ_k. By this construction,

$$\Phi_{k+1} = \Phi_k + W_{k+1}. \tag{1.5}$$

It is obvious that $\boldsymbol{\Phi} = \{\Phi_k : k \in \mathbb{Z}_+\}$ is a Markov chain, taking values in the real line $\mathbb{R} = (-\infty, \infty)$; the independence of the $\{W_k\}$ guarantees the Markovian nature of the chain $\boldsymbol{\Phi}$.

In this context, stability (as far as the gambling house is concerned) requires that $\boldsymbol{\Phi}$ eventually reaches $(-\infty, 0]$; a greater degree of stability is achieved from the same perspective if the time to reach $(-\infty, 0]$ has finite mean. Inevitably, of course, this stability is also the gambler's ruin.

Such a chain, defined by taking successive sums of i.i.d. random variables, provides a model for very many different systems, and is known as random walk.

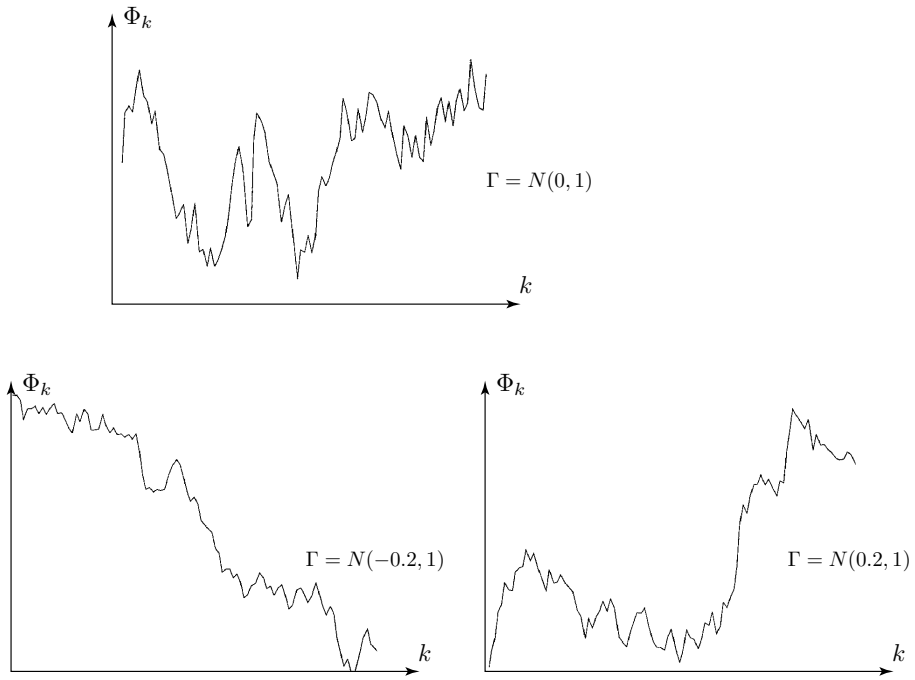

Figure 1.2: Random walk sample paths from three different models. The increment distributions is $\Gamma = N(0,1)$ for the path shown at top. The increment distribution is $\Gamma = N(-0.2,1)$ for the path shown on the lower left, and $\Gamma = N(+0.2,1)$ for the path shown on the lower right.

Random walk

Suppose that $\mathbf{\Phi} = \{\Phi_k; k \in \mathbb{Z}_+\}$ is a collection of random variables defined by choosing an arbitrary distribution for Φ_0 and setting for $k \in \mathbb{Z}_+$

(RW1)
$$\Phi_{k+1} = \Phi_k + W_{k+1}$$

where the W_k are i.i.d. random variables taking values in $\mathbb{R}$ with

$$\Gamma(-\infty, y] = \mathsf{P}(W_n \le y). \tag{1.6}$$

Then $\mathbf{\Phi}$ is called *random walk* on $\mathbb{R}$.

In Figure 1.2 we give sets of three sample paths of random walks with different distributions for Γ: all start at the same value but we choose for the winnings on each game