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 $\boldsymbol{\Phi} = \{\Phi_n : n \in T\}$, 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 $\boldsymbol{\Phi}$, 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 $\mathbb{Z}_+$

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

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 $\Phi_j, j \leq m$, 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 [147]). 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 [10]). 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 $X_{n-1}, \dots, X_{n-k}$, modified by the volatility of the market which is incorporated as a disturbance term W_n (see Krugman and Miller [142] 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 $\Phi_n = (X_n, \dots, X_{n-k+1})$, 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 [10]). 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 [220]). 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 [11]); more coarse analysis of large populations by time series models allows, as in (c), a Markovian representation (see Brockwell and Davis [32]); even the detailed and intricate cycle of the Canadian lynx seem to fit a Markovian model [188], [267]. 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, [211]). In particular, the calculation of posterior Bayesian distributions has been revolutionized through this route [244, 262, 264], 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 [75]): 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 [85]). 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 [222]). The spread of surnames may be modeled as Markovian (see [56]). The employment structure in a firm has a Markovian representation (see Bartholomew and Forbes [15]). 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 [127] for a very wide range of applications, both actual and potential, and Gray [89] 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 $\mathbf{Y} = \{Y_n\}$ 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

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

Merely by reformulating the model through defining the vectors

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

and setting $\boldsymbol{\Phi} = \{\Phi_n, n \geq 0\}$ (taking obvious care in defining $\{\Phi_0, \ldots, \Phi_{k-1}\}$), we can define from $\mathbf{Y}$ a Markov chain $\boldsymbol{\Phi}$. The motion in the first coordinate of $\boldsymbol{\Phi}$ reflects that of $\mathbf{Y}$, and in the other coordinates is trivial to identify, since Y_n becomes $Y_{(n+1)-1}$, and so forth; and hence $\mathbf{Y}$ 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" $\mathbf{x} = \{x_k,\ k \in \mathbb{Z}_+\}$ is defined inductively as

$$x_{k+1} = F x_k \tag{1.3}$$

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

Clearly, this is a multi-dimensional Markovian model: even if we know all of the values of $\{x_k, k \leq m\}$ 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.

In Figure 1.1 we show sample paths 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

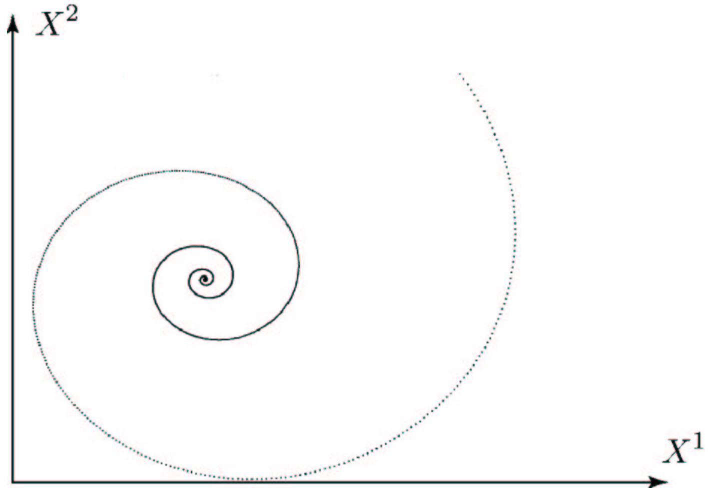

Fig. 1.1. Deterministic linear model on $\mathbb{R}^2$

instructive to realize that two very different types of behavior can follow from related choices of the matrix F. In Figure 1.1 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 $\mathbf{x} = \{x_k\}$ on $\mathbb{R}^n$ is defined inductively not only as a function of its past, but also of such a (deterministic) control sequence $\mathbf{u} = \{u_k\}$ 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 $x_j, j \leq k$ 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 $x_j, j \leq k$, and we describe this next.

Deterministic linear control model

Suppose $\mathbf{x} = \{x_k\}$ is a process on $\mathbb{R}^n$ and $\mathbf{u} = \{u_n\}$ 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 $\mathbf{x}$ 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 $\mathbf{X} = \{X_k\}$ 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 $\mathbf{X}$ is called the linear state space model driven by F, G, or the LSS(F,G) model, with *associated* control model LCM(F,G).

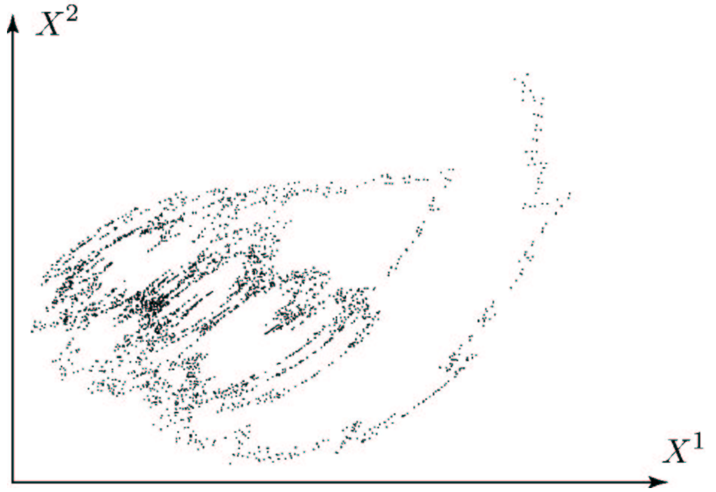

Fig. 1.2. Linear state space model on $\mathbb{R}^2$ with Gaussian noise

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. In Figure 1.2 we show sample paths corresponding to the choice of F as Figure 1.1 and $G = \binom{2.5}{2.5}$, with Γ taken as a bivariate Normal, or Gaussian, distribution $N(0, 1)$. This indicates that the addition of the noise variables $\mathbf{W}$ 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 [4, 39]. 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.

Random Walk on the Real Line

Suppose that $\boldsymbol{\Phi} = \{\Phi_k; k \in \mathbb{Z}_+\}$ 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 $\boldsymbol{\Phi}$ is called *random walk* on $\mathbb{R}$.

In Figure 1.3 , Figure 1.4 and Figure 1.5 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

(i) W having a Gaussian $\mathrm{N}(0,1)$ distribution, so the game is fair;

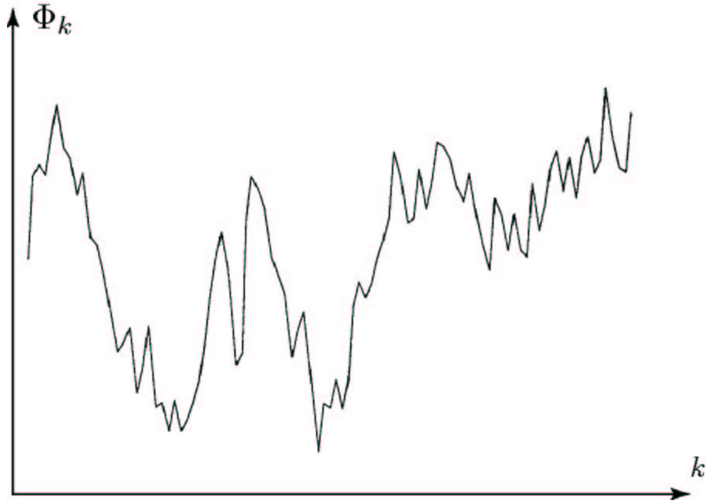

Fig. 1.3. Random walk paths with increment distribution $\Gamma = N(0,1)$

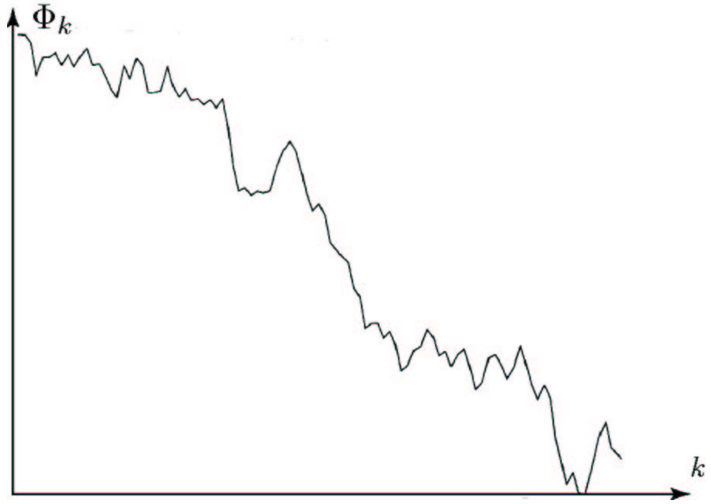

Fig. 1.4. Random walk paths with increment distribution $\Gamma = N(-0.2,1)$

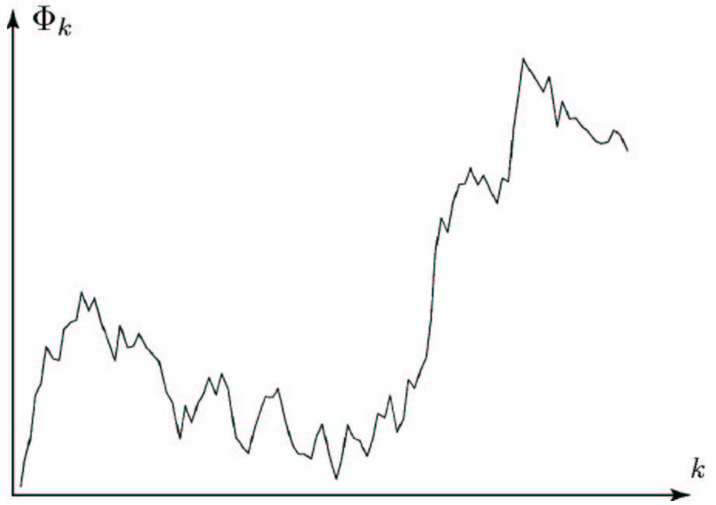

Fig. 1.5. Random walk paths with increment distribution $\Gamma = N(0.2, 1)$

(ii) W having a Gaussian N$(-0.2, 1)$ distribution, so the game is not fair, with the house winning one unit on average each five plays;

(iii) W having a Gaussian N$(0.2, 1)$ distribution, so the game modeled is, perhaps, one of "skill" where the player actually wins on average one unit per five games against the house.

The sample paths clearly indicate that ruin is rather more likely under case (ii) than under case (iii) or case (i): but when is ruin certain? And how long does it take if it is certain?

These are questions involving the stability of the random walk model, or at least that modification of the random walk which we now define.

Random walk on a half-line Although they come from different backgrounds, it is immediately obvious that the random walk defined by (RW1) is a particularly simple form of the linear state space model, in one dimension and with a trivial form of the matrix pair F, G in (LSS1). However, the models traditionally built on the random walk follow a somewhat different path than those which have their roots in deterministic linear systems theory.

Perhaps the most widely applied variation on the random walk model, which immediately moves away from a linear structure, is the random walk on a half-line.

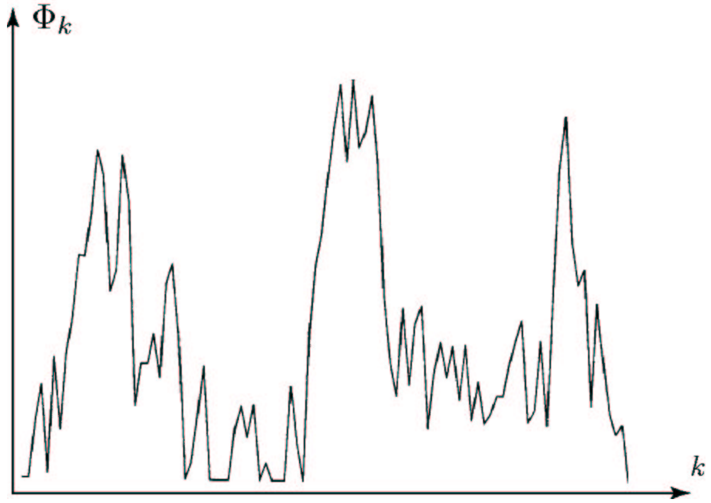

Fig. 1.6. Random walk paths stopped at zero, with increment distribution $\Gamma = N(-0.2, 1)$

Random Walk on a Half Line

Suppose $\boldsymbol{\Phi} = \{\Phi_k; k \in \mathbb{Z}_+\}$ is defined by choosing an arbitrary distribution for Φ_0 and taking

(RWHL1)

$$\Phi_{k+1} = [\Phi_k + W_{k+1}]^+ \tag{1.7}$$

where $[\Phi_k + W_{k+1}]^+ := \max(0, \Phi_k + W_{k+1})$ and again the W_k are i.i.d. random variables taking values in $\mathbb{R}$ with $\Gamma(-\infty, y] = \mathsf{P}(W \leq y)$.

Then $\boldsymbol{\Phi}$ is called *random walk on a half-line.*

This chain follows the paths of a random walk, but is held at zero when the underlying random walk becomes non-positive, leaving zero again only when the next positive value occurs in the sequence $\{W_k\}$.

In Figure 1.6 and Figure 1.7 we again give sets of sample paths of random walks on the half line $[0, \infty)$, corresponding to those of the unrestricted random walk in the previous section. The difference in the proportion of paths which hit, or return to, the state $\{0\}$ is again clear.

We shall see in Chapter 2 that random walk on a half line is both a model for storage systems and a model for queueing systems. For all such applications there

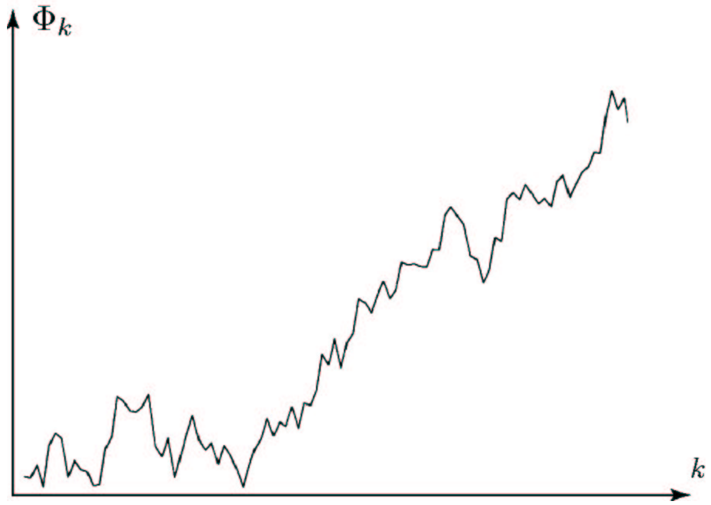

Fig. 1.7. Random walk paths stopped at zero, with increment distribution $\Gamma = N(+0.2, 1)$

are similar concerns and concepts of the structure and the stability of the models: we need to know whether a dam overflows, whether a queue ever empties, whether a computer network jams. In the next section we give a first heuristic description of the ways in which such stability questions might be formalized.

1.3 Stochastic Stability For Markov Models

What is "stability"?

It is a word with many meanings in many contexts. We have chosen to use it partly because of its very diffuseness and lack of technical meaning: in the stochastic process sense it is not well-defined, it is not constraining, and it will, we hope, serve to cover a range of similar but far from identical "stable" behaviors of the models we consider, most of which have (relatively) tightly defined technical meanings.

Stability is certainly a basic concept. In setting up models for real phenomena evolving in time, one ideally hopes to gain a detailed quantitative description of the evolution of the process based on the underlying assumptions incorporated in the model. Logically prior to such detailed analyses are those questions of the structure and stability of the model which require qualitative rather than quantitative answers, but which are equally fundamental to an understanding of the behavior of the model. This is clear even from the behavior of the sample paths of the models considered in the section above: as parameters change, sample paths vary from reasonably "stable" (in an intuitive sense) behavior, to quite "unstable" behavior, with processes taking larger or more widely fluctuating values as time progresses.

Investigation of specific models will, of course, often require quite specific tools: but the stability and the general structure of a model can in surprisingly wide-ranging circumstances be established from the concepts developed purely from the Markovian nature of the model.

We discuss in this section, again somewhat heuristically (or at least with minimal technicality: some "quotation-marked" terms will be properly defined later), various general stability concepts for Markov chains. Some of these are traditional in the Markov chain literature, and some we take from dynamical or stochastic systems theory, which is concerned with precisely these same questions under rather different conditions on the model structures.

1.3.1 Communication and recurrence as stability

We will systematically develop a series of increasingly strong levels of communication and recurrence behavior within the state space of a Markov chain, which provide one unified framework within which we can discuss stability.

To give an initial introduction, we need only the concept of the hitting time from a point to a set: let

$$\tau_A := \inf(n \geq 1 : \Phi_n \in A)$$

denote the first time a chain reaches the set A. This will be infinite for those paths where the set A is never reached.

In one sense the least restrictive form of stability we might require is that the chain does not in reality consist of two chains: that is, that the collection of sets which we can reach from different starting points is not different. This leads us to first define and study

(I) *φ-irreducibility* for a general space chain, which we approach by requiring that the space supports a measure φ with the property that for every starting point $x \in \mathsf{X}$

$$\varphi(A) > 0 \Rightarrow \mathsf{P}_x(\tau_A < \infty) > 0$$

where P_x denotes the probability of events conditional on the chain beginning with $\Phi_0 = x$.

This condition ensures that all "reasonable sized" sets, as measured by φ, can be reached from every possible starting point.

For a countable space chain φ-irreducibility is just the concept of irreducibility commonly used [40, 49], with φ taken as counting measure.

For a state space model φ-irreducibility is related to the idea that we are able to "steer" the system to every other state in $\mathbb{R}^n$. The linear control LCM(F,G) model is called *controllable* if for any initial states x_0 and any other $x^\star \in \mathsf{X}$, there exists $m \in \mathbb{Z}_+$ and a sequence of control variables $(u_1^\star, \dots u_m^\star) \in \mathbb{R}^p$ such that $x_m = x^\star$ when $(u_1, \dots u_m) = (u_1^\star, \dots u_m^\star)$. If this does not hold then for some starting points we are in one part of the space forever; from others we are in another part of the space. Controllability, and analogously irreducibility, preclude this.

Thus under irreducibility we do not have systems so unstable in their starting position that, given a small change of initial position, they might change so dramatically that they have no possibility of reaching the same set of states.

A study of the wide-ranging consequences of such an assumption of irreducibility will occupy much of Part I of this book: the definition above will be shown to produce remarkable solidity of behavior.

The next level of stability is a requirement, not only that there should be a possibility of reaching like states from unlike starting points, but that reaching such sets of states should be guaranteed eventually. This leads us to define and study concepts of

(II) *recurrence*, for which we might ask as a first step that there is a measure φ guaranteeing that for every starting point $x \in \mathsf{X}$

$$\varphi(A) > 0 \Rightarrow \mathsf{P}_x(\tau_A < \infty) = 1, \tag{1.8}$$

and then, as a further strengthening, that for every starting point $x \in \mathsf{X}$

$$\varphi(A) > 0 \Rightarrow \mathsf{E}_x[\tau_A] < \infty. \tag{1.9}$$

These conditions ensure that reasonable sized sets are reached with probability one, as in (1.8), or even in a finite mean time as in (1.9). Part II of this book is devoted to the study of such ideas, and to showing that for irreducible chains, even on a general state space, there are solidarity results which show that either such uniform (in x) stability properties hold, or the chain is unstable in a well-defined way: there is no middle ground, no "partially stable" behavior available.

For deterministic models, the recurrence concepts in (II) are obviously the same. For stochastic models they are definitely different. For "suitable" chains on spaces with appropriate topologies (the T-chains introduced in Chapter 6), the first will turn out to be entirely equivalent to requiring that "evanescence", defined by

$$\{\boldsymbol{\Phi} \to \infty\} = \bigcap_{n=0}^{\infty} \{\boldsymbol{\Phi} \in O_n \text{ infinitely often}\}^c \tag{1.10}$$

for a countable collection of open precompact sets $\{O_n\}$, has zero probability for all starting points; the second is similarly equivalent, for the same "suitable" chains, to requiring that for any $\varepsilon > 0$ and any x there is a compact set C such that

$$\liminf_{k\to\infty} P^k(x, C) \geq 1 - \varepsilon \tag{1.11}$$

which is *tightness* [24] of the transition probabilities of the chain.

All these conditions have the heuristic interpretation that the chain returns to the "center" of the space in a recurring way: when (1.9) holds then this recurrence is faster than if we only have (1.8), but in both cases the chain does not just drift off (or evanesce) away from the center of the state space.

In such circumstances we might hope to find, further, a long-term version of stability in terms of the convergence of the distributions of the chain as time goes by. This is the third level of stability we consider. We define and study

(III) the limiting, or *ergodic*, behavior of the chain: and it emerges that in the stronger recurrent situation described by (1.9) there is an "invariant regime" described by a measure π such that if the chain starts in this regime (that is, if Φ_0 has distribution π) then it remains in the regime, and moreover if the chain starts in some other regime then it converges in a strong probabilistic sense with π as a limiting distribution.

In Part III we largely confine ourselves to such ergodic chains, and find both theoretical and pragmatic results ensuring that a given chain is at this level of stability. For whilst the construction of solidarity results, as in Parts I and II, provides a vital underpinning to the use of Markov chain theory, it is the consequences of that stability, in the form of powerful ergodic results, that makes the concepts of very much more than academic interest.

Let us provide motivation for such endeavors by describing, with a little more formality, just how solid the solidarity results are, and how strong the consequent ergodic theorems are. We will show, in Chapter 13, the following:

Theorem 1.3.1 *The following four conditions are equivalent:*

(i) *The chain admits a unique probability measure π satisfying the invariant equations*

$$\pi(A) = \int \pi(dx)P(x,A), \qquad A \in \mathcal{B}(\mathsf{X}); \tag{1.12}$$

(ii) *There exists some "small" set $C \in \mathcal{B}(\mathsf{X})$ and $M_C < \infty$ such that*

$$\sup_{x \in C} \mathsf{E}_x[\tau_C] \leq M_C; \tag{1.13}$$

(iii) *There exists some "small" set C, some $b < \infty$ and some non-negative "test function" V, finite φ-almost everywhere, satisfying*

$$\int P(x,dy)V(y) \leq V(x) - 1 + b\mathbb{1}_C(x), \qquad x \in \mathsf{X}; \tag{1.14}$$

(iv) *There exists some "small" set $C \in \mathcal{B}(\mathsf{X})$ and some $P^\infty(C) > 0$ such that as $n \to \infty$*

$$\liminf_{n\to\infty} \sup_{x \in C} |P^n(x,C) - P^\infty(C)| = 0 \tag{1.15}$$

Any of these conditions implies, for "aperiodic" chains,

$$\sup_{A \in \mathcal{B}(\mathsf{X})} |P^n(x,A) - \pi(A)| \to 0, \qquad n \to \infty, \tag{1.16}$$

for every $x \in \mathsf{X}$ for which $V(x) < \infty$, where V is any function satisfying (1.14).

Thus "local recurrence" in terms of return times, as in (1.13) or "local convergence" as in (1.15) guarantees the uniform limits in (1.16); both are equivalent to the mere existence of the invariant probability measure π; and moreover we have in (1.14) an exact test based only on properties of P for checking stability of this type.

Each of (i)-(iv) is a type of stability: the beauty of this result lies in the fact that they are completely equivalent. Moreover, for this irreducible form of Markovian system, it is further possible in the "stable" situation of this theorem to develop asymptotic results, which ensure convergence not only of the distributions of the chain, but also of very general (and not necessarily bounded) functions of the chain (Chapter 14); to develop global rates of convergence to these limiting values (Chapter 15 and Chapter 16); and to link these to Laws of Large Numbers or Central Limit Theorems (Chapter 17).

Together with these consequents of stability, we also provide a systematic approach for establishing stability in specific models in order to utilize these concepts. The extension of the so-called "Foster-Lyapunov" criteria as in (1.14) to all aspects of stability, and application of these criteria in complex models, is a key feature of our approach to stochastic stability.

These concepts are largely classical in the theory of countable state space Markov chains. The extensions we give to general spaces, as described above, are neither so well-known nor, in some cases, previously known at all.

The heuristic discussion of this section will take considerable formal justification, but the end-product will be a rigorous approach to the stability and structure of Markov chains.

1.3.2 A dynamical system approach to stability

Just as there are a number of ways to come to specific models such as the random walk, there are other ways to approach stability, and the recurrence approach based on ideas from countable space stochastic models is merely one. Another such is through deterministic dynamical systems.

We now consider some traditional definitions of stability for a deterministic system, such as that described by the linear model (1.3) or the linear control model LCM(F,G).

One route is through the concepts of a *(semi) dynamical system*: this is a triple $(T, \mathcal{X}, d)$ where $(\mathcal{X}, d)$ is a metric space, and $T\colon \mathcal{X} \to \mathcal{X}$ is, typically, assumed to be continuous. A basic concern in dynamical systems is the structure of the *orbit* $\{T^k x : k \in \mathbb{Z}_+\}$, where $x \in \mathcal{X}$ is an *initial condition* so that $T^0 x := x$, and we define inductively $T^{k+1}x := T^k(Tx)$ for $k \geq 1$.

There are several possible dynamical systems associated with a given Markov chain.

The dynamical system which arises most naturally if X has sufficient structure is based directly on the transition probability operators P^k. If μ is an initial distribution for the chain (that is, if Φ_0 has distribution μ), one might look at the trajectory of distributions $\{\mu P^k : k \geq 0\}$, and consider this as a dynamical system $(P, \mathcal{M}, d)$ with $\mathcal{M}$ the space of Borel probability measures on a topological state space X, d a suitable metric on $\mathcal{M}$, and with the operator P defined as in (1.1) acting as $P\colon \mathcal{M} \to \mathcal{M}$ through the relation

$$\mu P(\,\cdot\,) = \int_{\mathsf{X}} \mu(dx) P(x,\,\cdot\,), \qquad \mu \in \mathcal{M}.$$

In this sense the Markov transition function P can be viewed as a deterministic map from $\mathcal{M}$ to itself, and P will induce such a dynamical system if it is suitably continuous. This interpretation can be achieved if the chain is on a suitably behaved space and has the *Feller* property that $Pf(x) := \int P(x, dy) f(y)$ is continuous for every bounded continuous f, and then d becomes a weak convergence metric (see Chapter 6).

As in the stronger recurrence ideas in (II) and (III) in Section 1.3.1, in discussing the stability of $\boldsymbol{\Phi}$, we are usually interested in the behavior of the terms P^k, $k \geq 0$, when k becomes large. Our hope is that this sequence will be bounded in some sense, or converge to some fixed probability $\pi \in \mathcal{M}$, as indeed it does in (1.16).

Four traditional formulations of stability for a dynamical system, which give a framework for such questions, are

(i) *Lagrange stability*: for each $x \in \mathcal{X}$, the orbit starting at x is a precompact subset of $\mathcal{X}$. For the system $(P, \mathcal{M}, d)$ with d the weak convergence metric, this is exactly tightness of the distributions of the chain, as defined in (1.11);

(ii) *Stability in the sense of Lyapunov*: for each initial condition $x \in \mathcal{X}$,

$$\lim_{y \to x} \sup_{k \geq 0} d(T^k y, T^k x) = 0,$$

where d denotes the metric on $\mathcal{X}$. This is again the requirement that the long term behavior of the system is not overly sensitive to a change in the initial conditions;

(iii) *Asymptotic stability*: there exists some fixed point x^* so that $T^k x^* = x^*$ for all k, with trajectories $\{x_k\}$ starting near x^* staying near and converging to x^* as $k \to \infty$. For the system $(P, \mathcal{M}, d)$ the existence of a fixed point is exactly equivalent to the existence of a solution to the invariant equations (1.12);

(iv) *Global asymptotic stability*: the system is stable in the sense of Lyapunov and for some fixed $x^* \in \mathcal{X}$ and every initial condition $x \in \mathcal{X}$,

$$\lim_{k \to \infty} d(T^k x, x^*) = 0. \tag{1.17}$$

This is comparable to the result of Theorem 1.3.1 for the dynamical system $(P, \mathcal{M}, d)$.

Lagrange stability requires that any limiting measure arising from the sequence $\{\mu P^k\}$ will be a probability measure, rather as in (1.16).

Stability in the sense of Lyapunov is most closely related to irreducibility, although rather than placing a global requirement on every initial condition in the state space, stability in the sense of Lyapunov only requires that two initial conditions which are sufficiently close will then have comparable long term behavior. Stability in the sense of Lyapunov says nothing about the actual boundedness of the orbit $\{T^k x\}$, since it is simply continuity of the maps $\{T^k\}$, uniformly in $k \geq 0$. An example of a system on $\mathbb{R}$ which is stable in the sense of Lyapunov is the simple recursion $x_{k+1} = x_k + 1$, $k \geq 0$. Although distinct trajectories stay close together if their initial conditions are similarly close, we would not consider this system stable in most other senses of the word.

The connections between the probabilistic recurrence approach and the dynamical systems approach become very strong in the case where the chain is both Feller and φ-irreducible, and when the irreducibility measure φ is related to the topology by the requirement that the support of φ contains an open set.

In this case, by combining the results of Chapter 6 and Chapter 18, we get for suitable spaces

Theorem 1.3.2 *For a φ-irreducible "aperiodic" Feller chain with* supp φ *containing an open set, the dynamical system $(P, \mathcal{M}, d)$ is globally asymptotically stable if and only if the distributions $\{P^k(x, \cdot)\}$ are tight as in (1.11); and then the uniform ergodic limit (1.16) holds.*

This result follows, not from dynamical systems theory, but by showing that such a chain satisfies the conditions of Theorem 1.3.1; these Feller chains are an especially useful subset of the "suitable" chains for which tightness is equivalent to the properties described in Theorem 1.3.1, and then, of course, (1.16) gives a result rather stronger than (1.17).

Embedding a Markov chain in a dynamical system through its transition probabilities does not bring much direct benefit, since results on dynamical systems in this level of generality are relatively weak. The approach does, however, give insights into ways of thinking of Markov chain stability, and a second heuristic to guide the types of results we should seek.

1.4 Commentary

This book does not address models where the time-set is continuous (when $\boldsymbol{\Phi}$ is usually called a Markov *process*), despite the sometimes close relationship between discrete and continuous time models: see Chung [49] or Anderson [5] for the classical countable space approach.

On general spaces in continuous time, there are a totally different set of questions that are often seen as central: these are exemplified in Sharpe [237], although the interested reader should also see Meyn and Tweedie [180, 181, 179] for recent results which are much closer in spirit to, and rely heavily on, the countable time approach followed in this book.

There has also been considerable recent work over the past two decades on the subject of more generally indexed Markov models (such as Markov *random fields*, where T is multi-dimensional), and these are also not in this book. In our development Markov chains always evolve through time as a scalar, discrete quantity.

The question of what to call a Markovian model, and whether to concentrate on the denumerability of the space or the time parameter in using the word "chain", seems to have been resolved in the direction we take here. Doob [68] and Chung [49] reserve the term chain for systems evolving on countable spaces with both discrete and continuous time parameters, but usage seems to be that it is the time-set that gives the "chaining". Revuz [223], in his Notes, gives excellent reasons for this.

The examples we begin with here are rather elementary, but equally they are completely basic, and represent the twin strands of application we will develop: the first, from deterministic to stochastic models via a "stochasticization" within the same functional framework has analogies with the approach of Stroock and Varadhan in their analysis of diffusion processes (see [260, 259, 102]), whilst the second, from basic independent random variables to sums and other functionals traces its roots back too far to be discussed here. Both these models are close to identical at this simple level. We give more diverse examples in Chapter 2.

We will typically use $\mathbf{X}$ and X_n to denote state space models, or their values at time n, in accordance with rather long established conventions. We will then typically use lower case letters to denote the values of related deterministic models. Regenerative models such as random walk are, on the other hand, typically denoted by the symbols $\boldsymbol{\Phi}$ and Φ_n, which we also use for generic chains.

The three concepts described in (I)-(III) may seem to give a rather limited number of possible versions of "stability". Indeed, in the various generalizations of deterministic dynamical systems theory to stochastic models which have been developed in the past three decades (see for example Kushner [149] or Khas'minskii [134]) there have been many other forms of stability considered. All of them are, however, qualitatively similar, and fall broadly within the regimes we describe, even though they differ in detail.

It will become apparent in the course of our development of the theory of irreducible chains that in fact, under fairly mild conditions, the number of different types of behavior is indeed limited to precisely those sketched above in (I)-(III). Our aim is to unify many of the partial approaches to stability and structural analysis, to indicate how they are in many cases equivalent, and to develop both criteria for stability to hold for individual models, and limit theorems indicating the value of achieving such stability.

With this rather optimistic statement, we move forward to consider some of the specific models whose structure we will elucidate as examples of our general results.