MARKOV CHAINS

ADAM M. JOHANSEN

1. Introduction

A Markov chain is, roughly speaking, some collection of random variables with a temporal ordering which have the property that *conditional upon the present, the future does not depend upon the past.* This concept, which can be viewed as a form of something known as the *Markov property*, will be made precise below, but the principle point is that such collections lie somewhere between one of *independent* random variables and a completely general collection which could be extremely complex to deal with.

Andrei Andreivich Markov commenced the analysis of such collections of random variables in 1907, and their analysis remains an active area of research to this day. The study of Markov chains is one of the great achievements of probability theory. In his seminal work [3], Andrei Nikolaevich Kolmogorov remarked "Historically, the independence of experiments and random variables represents the very mathematical concept that has given probability its peculiar stamp".

However, there are many situations in which it is necessary to consider sequences of random variables which *cannot* be considered to be independent. Kolmogorov went on to observe that "[Markov et al.] frequently fail to assume complete independence, they nevertheless reveal the importance of assuming analogous, weaker conditions, in order to obtain significant results". The aforementioned Markov property, the defining feature of the Markov chain, is such an *analogous, weaker condition* and it has proved both strong enough to allow many, powerful results to be obtained whilst weak enough to allow it to encompass a great many interesting cases.

Much of the development in probability theory during the latter part of the last century consisted of the study of sequences of random variables which are not entirely independent. Two weaker, but related conditions proved to be especially useful: the Markov property which defines the Markov chain and the *martingale* property. Loosely speaking, a martingale is a sequence of random variables whose expectation at any point in the future, conditional upon the past and present is equal to its present value. There is a broad and deep literature on the subject of martingales, which will not be discussed in this article. A great many people have worked on the theory of Markov chains, as well as their application to problems in a diverse range of areas, over the past century and it is not possible to enumerate them all here.

There are two principal reasons that Markov chains play such a prominent rôle in modern probability theory. The first is that they provide a powerful yet tractable framework in which to describe, characterise and analyse a broad class of sequences of random variables which find applications in numerous areas from particle transport through finite state machines and even in the theory of gene expression. The second is that a collection of powerful computational algorithms have been developed to provide samples from complicated probability distributions via the simulation of particular Markov chains: these *Markov chain Monte Carlo* methods are now ubiquitous in all fields in which it is necessary to obtain samples from complex probability distributions and this has driven much of the recent research in the field of Markov chains.

The areas in which Markov chains arise are far too numerous to list here, but here are some typical examples:

- Any collection of independent random variables forms a Markov chain: in this case, given the present the future is independent of the past *and the present.*
- The celebrated symmetric *random walk* over the integers provides a classical example: the next value taken by the chain is one more or less than the present value with equal probability, regardless of the route by which the present value was reached. Despite its simplicity, this example, and some simple generalisations can exhibit a great many interesting properties.
- Many popular board games have a Markov chain representation – for example, "Snakes and Ladders", in which there are 100 possible states for each counter (actually, there are somewhat fewer, as it is not possible to end a turn at the top of a snake or the bottom of a ladder) and the next state occupied by any particular counter is one of the six states which can be reached from the current one, each with equal probability. So, the next state is a function of the present state and an external, independent random variable corresponding to the roll of a die.
- More practically, the current amount of water held in a reservoir can be viewed as a Markov chain: the volume of water stored after a particular time interval will depend only upon the volume of water stored now and two random quantities: the amount of water leaving the reservoir and the amount of water entering the reservoir. More sophisticated variants of this model are used in numerous areas, particularly within the field of queueing theory (where water volume is replaced by customers awaiting service).
- The evolution of a finite state machine can be viewed as the evolution of a (usually deterministic) Markov chain.

It is usual to think of Markov chains as describing the trajectories of dynamic objects. In some circumstances there is a natural dynamic system associated with a collection of random variables with the right conditional independence structure – the random walk example discussed previously, for example, can be seen as moving from one position to the next, with the n^{th} element of the associated Markov chain corresponding to its position at discrete time index n. As the distribution of each random variable in the sequence depends only upon the value of the previous element of the sequence, one can endow any such collection (assuming that one can order the elements of the collection, which the definition of a Markov chain employed here ensures is always possible) with a dynamic structure. One simply views the distribution of each element, conditional upon the value of the previous one as being the probability of *moving* between those states at that time. This interpretation provides no great insight, but it can allow for simpler interpretations and descriptions of the behaviour of collections of random variables of the sort described here. Indeed, it is the image of a chain of states, each one leading to the next which suggests the term Markov chain.

2. Stochastic Processes

In order to proceed to the formal definition of a Markov chain, it is first necessary to make precise what is meant by a collection of random variables with some temporal ordering. Such a collection of random variables may be best characterised as a stochastic process. An E-valued *process* is a function $X : \mathcal{I} \to E$ which maps values in some index set $\mathcal{I}$ to some other space E. The evolution of the process is described by considering the variation of $X_i := X(i)$ with i. An E-valued *stochastic process* (or *random process*) can be viewed as a process in which, for each $i \in \mathcal{I}$, X_i is a random variable taking values in E.

Although a rich literature on more general situations exists, this article will consider only *discrete time stochastic processes* in which the index set $\mathcal{I}$ is the natural numbers, $\mathbb{N}$ (of course, any index set isomorphic to $\mathbb{N}$ can be used in the same framework by simple relabeling). The notation X_i is used to indicate the value of the process at time i (note that there need be no connection between the index set and *real* time, but this terminology is both convenient and standard). Note that the Markov property may be extended to continuous time processes in which the index set is the positive real numbers, and this leads to a collection of processes known as either *Markov processes* or *continuous time Markov chains*. Such processes are not considered further here, as they are of somewhat lesser importance in computer science and engineering applications. A rich literature on these processes does exist, and many of the results available in the discrete time case have continuous time analogues – indeed, some results may be obtained considerably more naturally in the continuous time setting.

At this point, a note on terminology is necessary. Originally, the term "Markov chain" was used to describe any stochastic process with the Markov property and a finite state space. Some references still use this definition today. However, in computer science, engineering and computational statistics it has become more usual to use the term to refer to any discrete time stochastic process with the Markov property, regardless of the state space, and this is the definition used here. Continuous time processes with the Markov property will be termed Markov processes, and little reference will be made to them. This usage is motivated by considerations arising from Markov chain Monte Carlo methods and is standard in more recent literature.

2.1. **Filtrations and Stopping Times.**

This section consists of some technical details which, whilst not essential to a basic understanding of the stochastic process or Markov chains in particular, are fundamental and will be encountered in any work dealing with these subjects.

A little more technical structure is generally required to deal with stochastic processes than with simple random variables. Whilst technical details are avoided as far as possible in this article, the following concept will be needed to understand much of the literature on Markov chains.

To deal with simple random variables, it suffices to consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ in which Ω is the set of events, $\mathcal{F}$ is the σ-algebra corresponding to the collection of measurable outcomes (i.e. the collection of subsets of Ω to which it is possible to assign a probability; typically the collection of all subsets of Ω in the discrete case) and $\mathbb{P}$ is the probability *measure*, which tells us the probability that any element of $\mathcal{F}$ contains the event which occurs: $\mathbb{P} : \mathcal{F} \to [0, 1]$. To deal with stochastic processes it is convenient to define a *filtered probability space* $(\Omega, \mathcal{F}, \{\mathcal{F}_i\}_{i\in\mathbb{N}}, \mathbb{P})$. The collection of sub-$\sigma$-algebras, $\{\mathcal{F}_i\}_{i\in\mathbb{N}}$, which is termed a *filtration*, has a particular structure:

$$\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}_n \subset \mathcal{F}_{n+1} \subset \cdots \subset \mathcal{F}$$

and its most important property is that, for any n, the collection of variables $X_1, X_2, \ldots, X_n$ must be measurable with respect to $\mathcal{F}_n$. Whilst much more generality is possible, it is usually sufficient to consider the *natural filtration* of a process: that is the one generated by the process itself. Given any collection of random variables of a common probability space, there exists a smallest σ-algebra with respect to which those random variables are jointly measurable. The natural filtration is the filtration generated by setting each $\mathcal{F}_n$ equal to the smallest σ-algebra with respect to which $X_1, \ldots, X_n$ are measurable. Only this filtration will be considered in the present article. An intuitive interpretation of this filtration, which provides increasingly fine sub-divisions of the probability space, is that $\mathcal{F}_n$ tells us how much information can be provided by knowledge of the values of the first n random variables: it tells us which events can be be distinguished given knowledge of $X_1, \ldots, X_n$.

It is natural when considering a process of this sort to ask questions about *random times*: is there anything to stop us defining additional random variables which have an interpretation as the index which identifies a particular time in the evolution of the process? In general, some care is required if these random times are to be useful: if the temporal structure is real, then it is necessary for us to be able to determine whether the time which has been reached so far is the time of interest, given some realisation of the process up to that time. Informally, one might require that $\{\tau = n\}$ can be ascribed a probability of zero or one, given knowledge of the first n states, for any n. In fact, this is a little stronger than the actual requirement, but it provides a simple interpretation which suffices for many purposes. Formally, if $\tau : \Omega \to \mathcal{I}$ is a random time, and the event $\{\omega : \tau(\omega) = n\} \in \mathcal{F}_n$ for all n, then τ is known as a *stopping time*. Note that this condition amounts to requiring that the event $\{\omega : \tau(\omega) = n\}$ is *independent* of all subsequent states of the chain, $X_{n+1}, X_{n+2}, \ldots$ conditional upon $X_1, \ldots, X_n$. The most common example of a stopping time is the *hitting time*, τ_A, of a set A:

$$\tau_A := \inf\{n : X_n \in A\},$$

which corresponds to the first time that the process enters the set A. Note that the apparently similar

$$\tau'_A = \inf\{n : X_{n+1} \in A\}$$

is *not* a stopping time (in any degree of generality) as the state of the chain at time $n + 1$ is not necessarily known in terms of the first t states.

Note that this distinction is not an artificial or frivolous one. Consider the chain produced by setting $X_n = X_{n-1} + W_n$ where $\{W_n\}$ are a collection of independent random variables corresponding to the value of a gambler's winnings, in dollars, in the n^{th} independent game which he plays. If $A = [10,000, \infty)$ then τ_A would correspond to the event of having won \$10,000 and, indeed, it would be possible to stop when this occurred. Conversely, if $A = (-\infty, -10,000]$, then τ'_A would correspond to the last time before that at which \$10,000 have been lost. Whilst many people would like to be able to stop betting immediately *before* losing money, it is not possible to know that one *will* lose the next one of a sequence of independent games.

Given a stopping time, τ, it is possible to define the *stopped process*, $X_1^\tau, X_2^\tau, \ldots$, associated with the process $X_1, X_2, \ldots$, which has the expected definition; writing $m \wedge n$ for the smaller of m and n, define $X_n^\tau = X_{\tau \wedge n}$. That is, the stopped process corresponds to the process itself at all times up to the random stopping time, after which it takes the value it had at that stopping time: it stops. In the case of τ_A, for example, the stopped process mirrors the original process until it enters A, and then retains the value it had upon entry to A for all subsequent times.

3. Markov Chains on Discrete State Spaces

Markov chains which take values in a discrete state space, such as the positive integers or the set of colours with elements red, green and blue, are relatively easy to define and to make use of. Note that this class of Markov chains includes those whose state space is countably infinite: as is often the case with probability little additional difficulty is introduced by the transition from finite to countable spaces, but considerably more care is needed to deal rigorously with uncountable spaces.

In order to specify the distribution a Markov chain on a discrete state space it is intuitively sufficient to provide an initial distribution, the marginal distribution of its first element, and the conditional distributions of each element given the previous one. In order to formalise this notion, and precisely what the *Markov property* referred to previously means, it is useful to consider the joint probability distribution of the first n elements of the Markov chain. Using the definition of conditional probability, it is possible to write the joint distribution of n random variables, $X_1, \ldots, X_n$, in the following form, using $X_{1:n}$ to denote the vector $(X_1, \ldots, X_n)$:

$$\mathbb{P}\left(X_{1:n} = x_{1:n}\right) = \mathbb{P}\left(X_1 = x_1\right) \prod_{i=2}^{n} \mathbb{P}\left(X_i = x_i | X_{1:i-1} = x_{1:i-1}\right).$$

The probability that each of the first n elements take particular values can be decomposed recursively as the probability that all but one of those elements takes the appropriate value and the conditional probability that the remaining element takes the specified value given that the other elements take the specified values.

This decomposition could be employed to describe the *finite dimensional distributions* (that is, the distribution of the random variables associated with finite subsets of $\mathcal{I}$) of any stochastic process. In the case of a Markov chain, the distribution of any element is influenced only by the previous state if the entire history is known: this is what is meant by the statement that "conditional upon the present, the future is independent of the past". This property may be written formally as

$$\mathbb{P}\left(X_n = x_n | X_{1:n-1} = x_{1:n-1}\right) = \mathbb{P}\left(X_n = x_n | X_{n-1} = x_{n-1}\right),$$

and so for any discrete state space Markov chain:

$$\mathbb{P}\left(X_{1:n} = x_{1:n}\right) = \mathbb{P}\left(X_1 = x_1\right) \prod_{i=2}^{n} \mathbb{P}\left(X_i = x_i | X_{i-1} = x_{i-1}\right).$$

As an aside, it is worthwhile to notice that Markov chains encompass a much broader class of stochastic processes than is immediately apparent. Given any stochastic process in which for all $n > L$ and $x_{1:n-1}$,

$$\mathbb{P}\left(X_n = x_n | X_{1:n-1} = x_{1:n-1}\right) = \mathbb{P}\left(X_n = x_n | X_{n-L:n-1} = x_{n-L:n-1}\right)$$

it suffices to consider a process Y on the larger space, E^L defined as

$$Y_n = (X_{n-L+1}, \ldots, X_n).$$

Note that $(X_{1-L}, \ldots, X_0)$ can be considered arbitrary without affecting the argument. Now, it is straightforward to determine that the distribution of Y_{n+1} depends only upon Y_n. In this way, any stochastic process with a finite memory may be cast into the form of a Markov chain on an extended space.

The Markov property, as introduced above, is more correctly known as the *weak Markov property*, and in the case of Markov chains in which the transition probability is not explicitly dependent upon the time index, it is normally written in terms of expectations of integrable test function $\xi : E^m \rightarrow \mathbb{R}$ where m may be any positive integer. The weak Markov property, in fact tells us that the expected value of the integral of any integrable test function over the next m states of a Markov chain depends only upon the value of the present state, so, for any n and any $x_{1:n}$:

$$\mathbb{E}\left[\xi(X_{n+1}, \ldots, X_{n+m}) | X_{1:n}\right] = \mathbb{E}\left[\xi(X_{n+1}, \ldots, X_{n+m+1}) | X_n\right].$$

It is natural to attempt to generalise this by considering random times, rather than deterministic ones. The *strong Markov property* requires that, for any stopping time τ, the following holds:

$$\mathbb{E}\left[\xi(X_{\tau+1},\ldots,X_{\tau+m})|X_{1:\tau}\right] = \mathbb{E}\left[\xi(X_{\tau+1},\ldots,X_{\tau+m+1})|X_{\tau}\right].$$

In continuous time settings, these two properties allow us to distinguish between weak and strong Markov processes (the latter is a strict subset of the former, because $\tau = n$ is a stopping time). However, in the discrete time setting, the weak and strong Markov properties are equivalent and are possessed by Markov chains as defined above.

It is conventional to view a Markov chain as describing the path of a dynamic object, which *moves* from one state to another as time passes. Many of the physical systems which can be described by Markov chains have precisely this property – for example, the motion of a particle in an absorbing medium. The position of the particle, together with an indication as to whether it has been absorbed or not may be described by a Markov chain whose states contain coordinates and an absorbed/not-absorbed flag. It is then natural to think of the initial state as having a particular distribution, say, $\mu(x_1) = \mathbb{P}(X_1 = x_1)$ and, furthermore, for there to be some *transition kernel* which describes the distribution of moves from a state x_{n-1} to a state x_n at time n, say, $K_n(x_{n-1}, x_n) = \mathbb{P}(X_n = x_n|X_{n-1} = x_{n-1})$. This allows us to write the distribution of the first n elements of the chain in the compact form:

$$\mathbb{P}(X_{1:n} = x_{1:n}) = \mu(x_1)\prod_{i=2}^{n} K_i(x_{i-1}, x_i).$$

There is nothing preventing these transition kernels from being explicitly dependent upon the time index, for example in the reservoir example presented above, one might expect both water usage and rainfall to have a substantial seasonal variation and so the volume of water stored tomorrow would be influenced by the date as well as the volume stored today. However, it is not surprising that for a great many systems of interest (and the vast majority of those used in computer simulation) that the transition kernel has no dependence upon the time. Markov chains which have the same transition kernel at all times are termed *time homogeneous* (or sometimes simply homogeneous) and will be the main focus of this article.

In the time homogeneous context, the n-step transition kernels denoted K^n which have the property that $\mathbb{P}(X_{m+n} = x_{m+n}|X_m = x_m) = K^n(x_m, x_{m+n})$ may be obtained inductively, as

$$K^n(x_m, x_{m+n}) = \sum_{x_{m+1}} K(x_m, x_{m+1})K^{n-1}(x_{m+1}, x_{m+n})$$

for any $n > 1$, whilst $K^1(x_m, x_{m+1}) = K(x_m, x_{m+1})$.

3.1. **A Matrix Representation.** The functional notation above is convenient, as it generalises to Markov chains on state spaces which are not discrete. However, discrete state space Markov chains exist in abundance in engineering and particularly in computer science. It is convenient to represent probability distributions on finite spaces as a row vector of probability values. To define such a vector, μ, simply set $\mu_i = \mathbb{P}(X = i)$ (where X is some random variable distributed according to μ). It is also possible to define a Markov kernel on this space by setting the elements of a matrix, K, equal to the probability of moving from a state i to a state j, i.e.:

$$K_{ij} = \mathbb{P}(X_n = j|X_{n-1} = i).$$

Whilst this may appear little more than a notational nicety, it has some properties which make manipulations particularly straightforward, for example, if $X_1 \sim \mu$

then:

$$\begin{aligned}\mathbb{P}\left(X_2 = j\right) &= \sum_i \mathbb{P}\left(X_1 = i\right)\mathbb{P}\left(X_2 = j|X_1 = i\right) \\ &= \sum_i \mu_i K_{ij} \\ &= \left(\mu K\right)_j,\end{aligned}$$

where μK denotes the usual vector matrix product and $(\mu K)_j$ denotes the j^{th} element of the resulting row vector. In fact, it can be shown inductively that $\mathbb{P}\left(X_n = j\right) = \left(\mu K^{n-1}\right)_j$, where K^{n-1} is the usual matrix power of K. Even more generally, the conditional distributions may be written in terms of the *transition matrix*, K:

$$\mathbb{P}\left(X_{n+m} = j|X_n = i\right) = (K^m)_{ij},$$

and so a great many calculations can be performed via simple matrix algebra.

3.2. **A Graphical Representation.** It is common to represent homogeneous, finite state space Markov chain graphically. A single directed graph with labeled edges suffices to completely describe the transition matrix of such a Markov chain. Together with the distribution of the initial state, this completely characterises the Markov chain. The vertices of the graph correspond to the states and those edges which exist illustrate the moves which it is possible to make. It is usual to label the edges with the probability associated with the move which they represent, unless all possible moves are equally probable.

A simple example, which also shows that the matrix representation can be rather difficult to interpret consists of the Markov chain obtained on the space $\{0, 1, \ldots, 9\}$ in which the next state is obtained by taking the number rolled on an unbiased die and adding it, modulo 10, to the present state *unless a 6 is rolled when the state is 9* in which case the chain retains its present value. This has a straightforward, but rather cumbersome matrix representation, in which:

$$K = \frac{1}{6}\begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

Figure 1 shows a graphical illustration of the same Markov transition kernel – transition probabilities are omitted in this case, as they are all equal. Whilst it may initially seem no simpler to interpret than the matrix, on closer inspection it becomes apparent that one can easily determine which states it is possible to reach any selected state from, which states it is possible to reach from it and which states it is possible to move between in a particular number of moves *without performing any calculations*. It is these properties which this representation make it very easy to interpret even in the case of Markov chains with large state spaces for which the matrix representation rapidly becomes very difficult to manipulate. Note the *loop* in the graph showing the possibility of remaining in state 9 – this is equivalent to the presence of a nonzero diagonal element in the transition matrix.

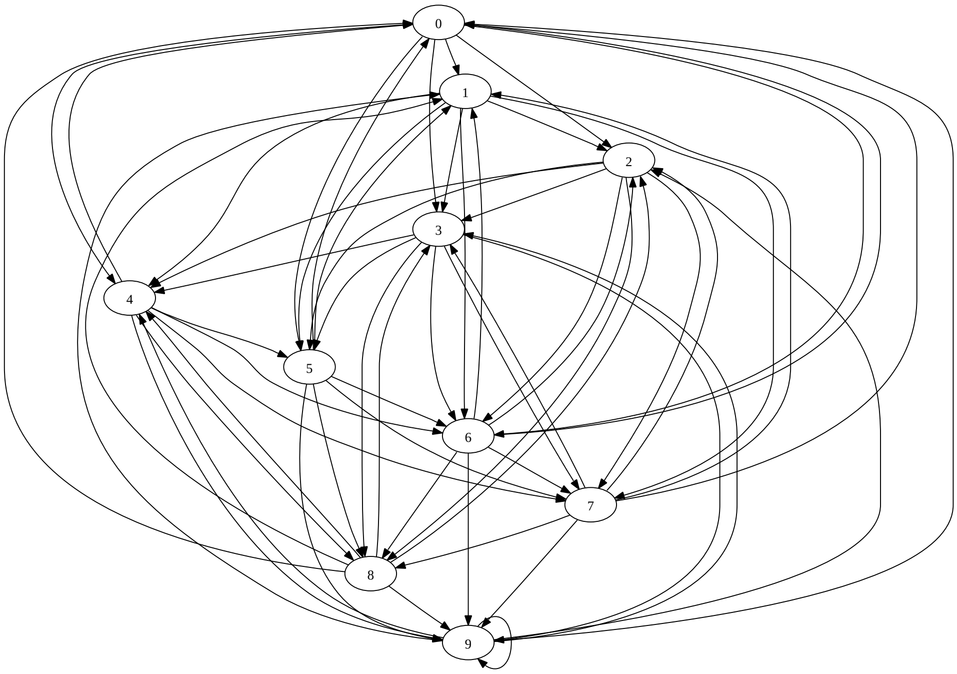

Figure 1. A graphical representation of a Markov chain.

4. Markov Chains on General State Spaces

In general, more subtle measure theoretic constructions are required to define or study Markov chains on *uncountable* state spaces – such as the real numbers, or the points in three dimensional space. In order to deal with fully general state space, a degree of measure theoretic probability beyond that which can be introduced in this article is required. Only Markov chains on some subset of d-dimensional, Euclidean space, $\mathbb{R}^d$ and to distributions and transition kernels which admit a density (for definiteness, with respect to Lebesgue measure – that which attributes to any interval mass corresponding to its length – over that space) will be considered here. $K_n(x, y)$ (or $K(x, y)$ in the time homogeneous case) denotes a density with the property that

$$\mathbb{P}\left(X_n \in A | X_{n-1} = x_{n-1}\right) = \int_A K_n(x_{n-1}, y)dy.$$

This approach has the great advantage that many concepts may be written for discrete and continuous state space cases in precisely the same manner, with the understanding that the notation refers to probabilities in the discrete case and densities in the continuous setting. In order to generalise things it is necessary to consider Lebesgue integrals with respect to the measures of interest, but essentially, one can replace equalities of densities with those of integrals over any measurable set and the definitions and results presented below will continue to hold. For a rigorous and concise introduction to general state space Markov chains, see [5] and for a much more detailed exposition [4] is highly recommended.

5. Stationary Distributions and Ergodicity

The ergodic hypothesis of statistical mechanics claims, loosely, that given a thermal system at equilibrium, the long term average occurrence of any given system configuration corresponds precisely to the average over an infinite ensemble of identically prepared systems at the same temperature. One area of great interest in the analysis of Markov chains is that of establishing conditions under which a (mathematically refined form of) this assertion can be shown to be true: when are averages obtained by considering those states occupied by a Markov chain over a long period of its evolution close to those which would be obtained by calculating the average under some distribution associated with that chain? Throughout this section integrals over the state space are used with the understanding that in the discrete case these integrals should be replaced by sums. This minimises the amount of duplication required to deal with both discrete and continuous state spaces, allowing the significant differences to be emphasised when they arise.

One of the most important properties of homogeneous Markov chains, particularly within the field of simulation, is that they can admit a *stationary* (or *invariant*) distribution. A transition kernel K is π-stationary if:

$$\int \pi(x)K(x,y)dx = \pi(y).$$

That is, given a sample $X = x$ from π, the distribution of a random variable, Y, drawn from $K(x,\cdot)$ is the same as that of X, although the two variables are, of course, not independent. In the discrete case this becomes:

$$\sum_i \pi(i)K(i,j) = \pi(j),$$

or, more succinctly, in the matrix representation, $\pi K = \pi$. The last of these reveals a convenient characterisation of the stationary distributions, where they exist, of a transition kernel: they are the left eigenvectors (or eigenfunctions in the general state space case) of the transition kernel with an associated eigenvalue of 1. Viewing the transition kernel as an operator on the space of distributions, the same interpretation is valid in the general state space case.

It is often of interest to simulate evolutions of Markov chains with particular stationary distributions. Doing so is the basis of Markov chain Monte Carlo methods, and is beyond the scope of this article. However, a number of theoretical concepts are required to determine when these distributions exist, when they are unique and when their existence is enough to ensure that a large enough sample path will have similar statistical properties to a collection of independent, identically distributed random variables from the stationary distribution. The remainder of this section is dedicated to the introduction of such concepts and the presentation of two results which are of great importance in this area.

One property useful in the construction of Markov chains with a particular invariant distributions is that of *reversibility.* A stochastic process is termed reversible if the statistical properties of its time-reversal are the same as those of the process itself. To make this concept more formal, it is useful to cast things in terms of certain joint probability distributions. A stationary process is reversible, if for any n, m the following equality holds for all measurable sets $A_n, \dots, A_{n+m}$:

$$\mathbb{P}\left(X_n \in A_n, \dots X_{n+m} \in A_{n+m}\right) = \mathbb{P}\left(X_n \in A_n, \dots, X_{n-m} \in A_{n+m}\right).$$

It is simple to verify that, in the context of a Markov chain, this is equivalent to the *detailed balance* condition:

$$\mathbb{P}\left(X_n \in A_n, X_{n+1} \in A_{n+1}\right) = \mathbb{P}\left(X_n \in A_{n+1}, X_{n+1} \in A_n\right).$$

A Markov chain with kernel K is said to satisfy detailed balance for a distribution π if:

$$\pi(x)K(x,y) = \pi(y)K(y,x).$$

It is straightforward to verify that, if K is π-reversible, then π is a stationary distribution of K:

$$\begin{aligned}\int \pi(x)K(x,y)dy &= \int \pi(y)K(y,x)dy \\ \pi(x) &= \int \pi(y)K(y,x)dy.\end{aligned}$$

This is particularly useful, as the detailed balance condition is straightforward to verify.

Given a Markov chain with a particular stationary distribution, it is important to be able to determine whether, over a long enough period of time, the chain will explore all of the space which has positive probability under that distribution. This leads us to concepts of *accessibility, communication structure* and *irreducibility.*

In the discrete case, a state, j, is said to be *accessible* from another state, i, written as $i \rightarrow j$, if for some n, $K^n(i,j) > 0$. That is, a state which is accessible from some starting point, is one which can be reached with positive probability in some number of steps. If i is accessible from j and j is also accessible from i, then the two states are said to *communicate* and this is written as $i \leftrightarrow j$. Given the Markov chain on the space $E = \{0,1,2\}$ with transition matrix:

$$K = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix},$$

it is not difficult to verify that the uniform distribution $\mu = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ is invariant under the action of K. However, if $X_1 = 0$ then $X_n = 0$ for all n: the chain will never reach either of the other states, whilst started from $X_1 \in \{1,2\}$ the chain will never reach 0. This chain is *reducible*: there are disjoint regions of the state space which do not communicate. Furthermore, it has multiple stationary distributions, $(1,0,0)$ and $(0, \frac{1}{2}, \frac{1}{2})$ are both invariant under the action of K. In the discrete setting, a chain is *irreducible* if all states communicate: starting from any point in the state space, any other point may be reached with positive probability in some finite number of steps.

Whilst these concepts are adequate for dealing with discrete state spaces, a little more subtlety is required in more general settings: as ever, when dealing with probability on continuous spaces, the probability associated with individual states is generally zero and it is necessary to consider integrals over finite regions. The property which is captured by *irreducibility* is that, wherever the chain starts from, there is a positive probability of it reaching anywhere in the space. In order to generalise this to continuous state spaces it suffices to reduce the strength of this statement very slightly to: from "most" starting points, the chain has a positive probability of reaching any region of the space which itself has positive probability. To make this precise, a Markov chain of stationary distribution π is said to be π-irreducible if, for all x (except for those lying in a set of exceptional points which has probability 0 under π), and all sets A with the property that $\int_A \pi(x)dx > 0$,

$$\exists n : \int_A K^n(x,y)dy > 0.$$

The terms *strongly irreducible* and *strongly π-irreducible* are sometimes used when the irreducibility or π-irreducibility condition, respectively, holds for $n = 1$. Notice that any irreducible Markov chain is π-irreducible with respect to any measure π.

These concepts allow us to determine whether a Markov chain has a "joined-up" state space: whether it is possible to move around the entire space (or at least that part of the space which has mass under π). However, it tells us nothing about *when* it is possible to reach these points. Consider the difference between the following two transition matrices on the space $\{0, 1\}$, for example:

$$K = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} \text{ and } L = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Both admit $\pi = (\frac{1}{2}, \frac{1}{2})$ as a stationary distribution, and both are irreducible. However, consider their respective marginal distributions after a number of iterations:

$$K^n = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} \text{ whilst } L^n = \begin{cases} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} & n \text{ odd} \\ \\ \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} & n \text{ even} \end{cases}.$$

In other words, if $\mu = (\mu_1, \mu_2)$, then the Markov chain associated with K has distribution $\mu K^n = (\frac{1}{2}, \frac{1}{2})$ after n iterations, whilst that associated with L has distribution (μ_2, μ_1) after any odd number of iterations and distribution (μ_1, μ_2) after any even number. L, then, never forgets its initial conditions and it is *periodic*.

Whilst this is a rather contrived example, it is clear that such periodic behaviour is significant and a precise characterisation is needed. This is straightforward in the case of discrete state space Markov chains. For any state in the space, i, its *period* is defined, using gcd to refer to the greatest common divisor (i.e. the largest common factor) as:

$$d = \gcd \{n : K^n(i, i) > 0\}.$$

Thus, in the case of L, above, both states have a period $d = 2$. In fact, it can easily be shown, that any pair of states which communicate must have the same period. Thus, irreducible Markov chains have a single period, 2, in the case of L, above and 1, in the case of K. Irreducible Markov chains may be said to have a period themselves, and when this period is 1, they are termed *aperiodic*.

Again, more subtlety is required in the general case. It is clear that something is needed to fill the rôle that individual states play in the discrete state space case, and that individual states are not appropriate in the continuous case. A set of events which is small enough that it is, in some sense, homogeneous and large enough that it has positive probability under the stationary distribution is required. A set C is termed *small* if there exists some integer n, some probability distribution ν and some $\epsilon > 0$ such that the following condition holds:

$$\inf_{x \in C} K^n(x, y) \geq \epsilon \nu(y).$$

This tells us that for any point in C, with probability ϵ, the distribution of the next state the chain enters is independent of *where in C* it is. In that sense C is small, and these sets are precisely what is necessary to extend much of the theory of Markov chains from the discrete state space case to a more general setting. In particular, it is now possible to extend the notion of period from the discrete state space setting to a more general one. Note that in the case of irreducible Markov chains on a discrete state space, the entire state space is small.

A Markov chain has a cycle of length d if there exists a small set C such that the greatest common divisor of the length of paths from C to a measurable set of positive probability B is d. If the largest cycle possessed by a Markov chain has length 1, then that chain is *aperiodic*. In the case of π-irreducible chains, every state has a common period (except a set of events of probability 0 under π), and the above definition is equivalent to the more intuitive (but more difficult to verify) condition, that there exists a partition of the state space, E, into d disjoint subsets $E_1, \ldots, E_d$ with the property that $\mathbb{P}(X_{n+1} \notin E_j | X_n \in E_i) = 0$ if $j = i + 1 \bmod d$.

Thus far, concepts which allow us to characterise those Markov chains which can reach every important part of the space, and which exhibit no periodic structure have been introduced. Nothing has been said about *how often* a given region of the space might be visited. This is a particularly important point: there is a qualitative difference between chains which have a positive probability of returning to a set infinitely often, and those which can only visit it finitely many times. Let η_A denote the number of times that a set A is visited by a Markov chain, that is, $\eta_A = |\{X_n \in A : n \in \mathbb{N}\}|$. A π-irreducible Markov chain is recurrent if $\mathbb{E}\left[\eta_A\right] = \infty$ for every A with positive probability under π. Thus a recurrent Markov chain is one with positive probability of visiting any significant (with respect to π) part of the state space infinitely often: it does not always escape to infinity. A slightly stronger condition is termed *Harris recurrence*, it requires that every significant state *is* visited infinitely often (rather than this event having positive probability), i.e. $\mathbb{P}\left(\eta_A = \infty\right) = 1$ for every set A for which $\int_A \pi(x)dx > 0$. A Markov chain which is not recurrent is termed *transient*.

The following example illustrates the problems which can arise if a Markov chain is π-recurrent but *not* Harris recurrent. Consider the Markov chain over the positive integers with the transition kernel defined by:

$$K(x,y) = x^{-2}\delta_1(y) + (1-x^{-2})\delta_{x+1}(y),$$

where for any state, x, δ_x denotes the probability distribution which places all of its mass at x. This kernel is clearly δ_1-recurrent: if the chain is started from 1, it stays there deterministically. However, as the sum

$$\sum_{k=2}^{\infty} \frac{1}{k^2} < \infty,$$

the Borel-Cantelli lemma ensures that whenever the chain is started for any x greater than 1, there is positive probability that the chain will *never* visit state 1 – the chain is π-recurrent, but it is not Harris recurrent. Although this example is somewhat contrived, it illustrates an important phenomenon – and one which often cannot easily be detected in more sophisticated situations. It has been suggested that Harris recurrence can be interpreted as a guarantee that their exist no such pathological system trajectories: there are no parts of the space from which the chain will escape to infinity rather than returning to the support of the stationary distribution.

It is common to refer to a π-irreducible, aperiodic, recurrent Markov chain as being ergodic, and an ergodic Markov chain which is also Harris recurrent as being *Harris ergodic*. These properties suffice to ensure that the Markov chain will, on average visit every part of the state space in proportion to its probability under π, that it exhibits no periodic behaviour in doing so and that it might (or will, in the Harris case) visit regions of the state space with positive probability infinitely often. Actually, ergodicity tells us that a Markov chain eventually forgets its initial conditions – after a sufficiently long time has elapsed, the current state provides arbitrarily little information about the initial state. There are a number of stronger forms of ergodicity which provide information about the *rate* at which the initial conditions are forgotten, these are covered in great detail by [4]. Intuitively, if a sequence of random variables forgets where it has been, but has some stationary distribution then one would expect the distribution of sufficiently widely separated samples to approximate that of independent samples from that stationary distribution. This intuition can be made rigorous, and is strong enough to tell us quite a lot about the distribution of large samples obtained by iterative application of the Markov kernel and the sense in which approximations of integrals obtained by using the empirical average obtained by taking samples from the chain might converge to their integral under the stationary measure. This section is concluded with two of the most important results in the theory of Markov chains.

The *ergodic theorem* provides an analogue of the law of large numbers for independent random variables: it tells us that under suitable regularity conditions, the averages obtained from the sample path of a Markov chain will converge to the expectation under the stationary distribution of the transition kernel. This is the mathematically refined, rigorously proved form of the ergodic hypothesis alluded to at the start of this section. There are a great many variants of this theorem, one particularly simple form is the following: if $\{X_n\}$ is a Harris ergodic Markov chain of invariant distribution π, then the following strong law of large numbers holds for any π-integrable function $f : E \to \mathbb{R}$ (convergence is with probability one):

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f(X_i) \to \int f(x)\pi(x)dx.$$

This is a particular case of [6, p. 241, Theorem 6.63], and a proof of the general theorem is given there. The same theorem is also presented with proof in [4, p. 433, Theorem 17.3.2].

A *central limit theorem* also exists, and tells us something about the rate of convergence of averages under the sample path of the Markov chain. Under technical regularity conditions (see [2] for a summary of various combinations of conditions) it is possible to obtain a central limit theorem for the ergodic averages of a Harris recurrent, μ-invariant Markov chain, and a function which has at least two finite moments, $f : E \to \mathbb{R}$ (with $\mathbb{E}[f] < \infty$ and $\mathbb{E}\left[f^2\right] < \infty$)[1].

$$\lim_{n\to\infty} \sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \int f(x)\pi(x)dx\right] \xrightarrow{d} \mathcal{N}(0, \sigma^2(f)),$$

$$\sigma^2(f) = \mathbb{E}\left[(f(X_1) - \bar{f})^2\right] + 2\sum_{k=2}^{\infty} \mathbb{E}\left[(f(X_1) - \bar{f}\,)(f(X_k) - \bar{f}\,)\right]$$

where $\xrightarrow{d}$ denotes convergence in distribution, $\mathcal{N}(0, \sigma^2)$ is the normal distribution of mean 0 and variance $\sigma^2(f)$ and $\bar{f} = \int f(x)\pi(x)dx$.

A great many refinements of these results exist in the literature. In particular, cases in which conditions may be relaxed or stronger results proved have been very widely studied. It is of particular interest in many cases to obtain quantitative bounds on the rate of convergence of ergodic averages to the integral under the stationary distribution. In general, it is very difficult to obtain meaningful bounds of this sort for systems of real practical interest, although some progress has been made in recent years.

6. Selected Extensions and Related Areas

It is unsurprising that a field as successful as that of Markov chains has a number of interesting extensions, and related areas. This section briefly describes two of these.

So-called *adaptive Markov chains* have received a significant amount of attention in the field of Monte Carlo methodology in recent years. These are systems in which the transition kernel used at each iteration is adjusted depending upon the entire history of the system or some statistical summary of that history. Whilst these adaptive systems are attractive from a practical viewpoint, as they allow for automatic tuning of parameters and promise simpler implementation of Monte

[1] Depending upon the combination of regularity conditions assumed, it may be necessary to have a finite moment of order $2 + \delta$

Carlo methods in the future, a great deal of care must be taken when analysing them. It is important to notice that, because the transition kernel depends upon more than the current state at the time of its application, it does not give rise to a Markov chain.

Feynman-Kac formulae were first studied in the context of describing physical particle motion. They describe a sequence of probability distributions obtained from a collection of Markov transition kernels, M_n, and a collection of potential functions, G_n. Given a distribution η_{n-1} at time $n-1$, the system is *mutated* according to the transition kernel to produce an updated distribution,

$$\hat{\eta}_n(x_n) = \int \eta_{n-1}(x_{n-1}) M_n(x_{n-1}, x_n) dx_{n-1},$$

before weighting the probability of each state / region of the space according to the value of the potential function:

$$\eta_n(x_n) = \frac{\hat{\eta}_n(x_n) G_n(x_n)}{\int \hat{\eta}_n(x) G_n(x) dx}.$$

There are a number of convenient ways of interpreting such sequences of distributions. One is that if η_{n-1} describes the distribution of a collection of particles at time $n-1$, which have dynamics described by the Markov kernel M_n in an absorbing medium which is described by the potential function G_n (in the sense that the smaller the value of G_n at a point, the greater the probability that a particle at that location is absorbed) then η_n describes the distribution of those particles which have not been absorbed at time n. These systems have found a great deal of application in the fields of Monte Carlo methodology, particularly sequential and population-based methods, and genetic algorithms. The latter gives rise to another interpretation: the Markov kernel can be seen as describing the mutation undergone by individuals within a population, and the potential function $G_n(x_n)$ the fitness of an individual with genetic make-up x_n which governs the success of that individual in a selection step.

Alternatively, one can view the evolution of a Feynman-Kac system as a *non-linear Markov Chain* in which the distribution of X_n depends upon both X_{n-1} and its distribution, η_{n-1}. That is, if $X_{n-1} \sim \eta_{n-1}$ then the distribution of X_n is given by

$$\eta_n(\cdot) = \int \eta_{n-1}(x_{n-1}) K_{n,\eta_{n-1}}(x_{n-1}, \cdot) dx_{n-1}$$

, where the non-linear Markov Kernel K_{n,η_n} is usually defined as the composition of selection and mutation steps (numerous such kernels may be associated with any particular Feynman-Kac flow).

An excellent monograph on Feynman-Kac formulae and their mean field approximations has recently been written [1].

References

[1] P. Del Moral. *Feynman-Kac formulae: genealogical and interacting particle systems with applications*. Probability and Its Applications. Springer Verlag, New York, 2004.

[2] G. L. Jones. On the Markov chain central limit theorem. *Probability Surveys*, 1:299–320, 2004.

[3] A. N. Kolmogorov. *Foundations of the Theory of Probability*. Chelsea Publishing Company, second english edition, 1956.

[4] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Springer Verlag, 1993.

[5] E. Nummelin. *General Irreducible Markov Chains and Non-Negative Operators*. Number 83 in Cambridge Tracts in Mathematics. Cambridge University Press, 1st paperback edition, 1984.

[6] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer Verlag, New York, second edition, 2004.