Advanced Simulation Methods

Chapter 4 - Elements of Markov chains Theory

We present a very brief and elementary introduction to the theory of Markov chains in order to provide some justification of the Markov chain Monte Carlo methods presented in this course. A rigorous treatment of this theory requires measure theoretic concepts which are beyond the scope of this course. We will neglect these important issues here and try to preserve the main ideas. Hence if you have no knowledge of measure theory, just ignore the references to measurability.

1 Discrete State Space Markov Chains

1.1 Stochastic Processes

A discrete-time $\mathbb{X}$-valued stochastic process is a process where, for each $t \in \mathbb{N}$, X_t is a random variable taking values in a space $\mathbb{X}$. Typically we will deal with either discrete spaces (such as a finite set like $\{1, 2, \ldots, d\}$ for some $d \in \mathbb{N}$ or a countable set like the set of integers $\mathbb{Z}$), or continuous spaces (such as $\mathbb{R}$ or $\mathbb{R}^d$ for some $d \in \mathbb{N}$). The space $\mathbb{X}$ is often called the state space. In order to characterise a discrete-time stochastic process, it is sufficient to know all of its finite dimensional distributions, that is, the joint distributions of the process at any collection of finitely many times. For a collection of times $(t_1, \ldots, t_n)$ and a collection of measurable sets of $\mathbb{X}$, $(A_{t_1}, ..., A_{t_n})$, the process is associated with the joint probability

$$\mathbb{P}\left(X_{t_1} \in A_{t_1}, X_{i_2} \in A_{i_2}, ..., X_{t_n} \in A_{t_n}\right).$$

The fact that those probabilities completely define a stochastic process comes from an important result called the Kolmogorov extension theorem. Note that remarkably, it allows to define a process $(X_t)_{t\in\mathbb{N}}$ which is an infinite-dimensional object (because of the "$t \in \mathbb{N}$") using only finite objects (like the joint probability above), under some consistency conditions omitted here for simplicity. To define a stochastic process, all we need to do is to specify these finite dimensional distributions. We will focus here on the class of Markov processes which are both simple and the most useful class of stochastic processes in the context of Monte Carlo methods. We will see that their specification only requires an initial distribution and a transition probability.

Let us first consider the case of discrete state spaces, i.e. $|\mathbb{X}|$ is finite or countably infinite. We can assume, without loss of generality, that the state space $\mathbb{X}$ is $\mathbb{N}$. In this context, we can work with the actual probability of the process having a particular value at any time (in the case of continuous random variables admitting a probability density function with respect to the Lebesgue measure then the probability of taking any particular value is zero). For any $t \in \mathbb{N}$, we always have the following decomposition, for a collection of points $(x_1, \ldots, x_t)$ in $\mathbb{X}$,

$$\begin{aligned}
&\mathbb{P}(X_1 = x_1, X_2 = x_2, ..., X_t = x_t) \qquad (1)\\
&\quad = \mathbb{P}(X_1 = x_1, X_2 = x_2, ..., X_{t-1} = x_{t-1})\mathbb{P}(X_t = x_t | X_1 = x_1, X_2 = x_2, ..., X_{t-1} = x_{t-1})\\
&\quad = \mathbb{P}(X_1 = x_1)\prod_{s=2}^{t}\mathbb{P}(X_s = x_s | X_1 = x_1, ..., X_{s-1} = x_{s-1}).
\end{aligned}$$

From this decomposition, we can construct all of the finite dimensional distributions using simply the sum and product rules of probability. In many situations, we can assume that the distribution of X_s given its "past" $(X_1, \ldots, X_{s-1})$ depends only upon X_{s-1} and is otherwise independent of the other values it took to reach X_{s-1}; i.e. we have

$$\mathbb{P}(X_s = x_s | X_1 = x_1, ..., X_{s-1} = x_{s-1}) = \mathbb{P}(X_s = x_s | X_{s-1} = x_{s-1}). \qquad (2)$$

Stochastic processes for which (2) holds are known as discrete time Markov processes or simply as Markov chains in the Monte Carlo literature.

When dealing with discrete state spaces, it is often convenient to associate a row vector with any probability distribution. Now, given a random variable X on $\mathbb{X}$, we say that X has distribution μ for some vector μ with the notation:

$$\forall x \in \mathbb{X} \quad \mathbb{P}(X = x) \equiv \mu(x).$$

1.2 Homogeneous Markov Chains

Homogeneous Markov chains are Markov processes with conditional probabilities that do not depend on the time index, i.e.

$$\forall x, y \in \mathbb{X} \quad \forall t, s \in \mathbb{N} \quad \mathbb{P}(X_t = y | X_{t-1} = x) = \mathbb{P}(X_{t+s} = y | X_{t+s-1} = x). \tag{3}$$

In this setting, we can introduce the associate transition matrix $K(i,j) = K_{ij} = \mathbb{P}(X_t = j | X_{t-1} = i)$. K is often referred to as the kernel of the Markov chain. If we call μ_t the distribution of X_t, $\mu_t(i) = \mathbb{P}(X_t = i)$, then it follows by combining (1)-(2)-(3) that the joint distribution of the chain over any finite time horizon satisfies

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, ..., X_t = x_t) = \mu_1(x_1) \prod_{s=2}^{t} K_{x_{s-1} x_s}.$$

We can also define K^n with entries $K^n(i,j)$, the matrix of transition from i to j in n steps:

$$K^n_{ij} = \mathbb{P}(X_{t+n} = j | X_t = i).$$

We obtain for any i, j the so-called Chapman-Kolmogorov equation

$$K^{m+n}_{ij} = \sum_k K^m_{ik} K^n_{kj}.$$

Indeed for any i, j

$$\begin{aligned}
K^{m+n}_{ij} &= \mathbb{P}(X_{m+n} = j | X_0 = i) \\
&= \sum_k \mathbb{P}(X_{m+n} = j, X_m = k | X_0 = i) \\
&= \sum_k \mathbb{P}(X_{m+n} = j | X_m = k, X_1 = i) \mathbb{P}(X_m = k | X_0 = i) \\
&= \sum_k \mathbb{P}(X_{m+n} = j | X_m = k) \mathbb{P}(X_m = k | X_0 = i) \\
&= \sum_k K^n_{kj} K^m_{ik}.
\end{aligned}$$

Chapman-Kolmogorov equation implies that indeed K^n is the n^{th} matrix power of K, and hence the notation is consistent with standard linear algebra. In terms of the marginal laws of X_t we obtain the expression

$$\mu_{t+1}(j) = \sum_i \mu_t(i) K_{ij}.$$

If $\mathbb{X}$ is finite this is nothing else than a standard vector-matrix multiplication, hence we write the equation simply as

$$\mu_{t+1} = \mu_t K. \tag{4}$$

Similarly we obtain $\mu_{t+n} = \mu_t K^n$.

To summarise, homogeneous Markov chains can be characterised as follows. First the distribution μ_0 of X_0 must be specified. Then a transition kernel K that specifies the law of X_t given X_{t-1} at any time t. The distribution μ_0 and the transition K completely define the Markov chain (X_t), using the Chapman-Kolmogorov equation above and the fact that finite-dimensional joint distributions characterise stochastic processes.

Although homogeneous Markov chains are predominantly used in Monte Carlo, there are also popular techniques such as simulated annealing, adaptive Markov chain Monte Carlo and particle filters which rely on non-homogeneous Markov chains. These processes are more complex to analyse and we will restrict ourselves to homogeneous chains henceforth.

1.3 Important Properties

1.3.1 Irreducibility

We review here some of the main concept/properties associated to Markov chains. We first consider how states communicate to each other under a given Markov chain transition kernel.

Definition 1 ***(Accessibility)***. *A state y is accessible from a state x, written as $x \to y$ if, for a discrete state space Markov chain,* $\inf\{t : \mathbb{P}(X_t = y | X_1 = x) > 0\} < \infty$. *This can be rewritten equivalently as* $\inf\left\{t : K_{xy}^t > 0\right\} < \infty$.

In layman's terms, $x \to y$ means that starting from x there is a positive probability of reaching y at some finite time in the future when we "move" according to the Markov kernel K. It is now useful to consider cases in which one can traverse the entire space, or some subset of it, starting from any point.

Definition 2 ***(Communication)***. *Two states $x, y \in \mathbb{X}$ are said to communicate if and only if $x \to y$ and $y \to x$.*

These notions allow us to characterise the so-called communication structure of the associated Markov chain to some degree, noting from which points it is possible to travel both to and back from. We now introduce a concept to describe the properties of the full state space, or significant parts of it, rather than individual states.

Definition 3 ***(Irreducibility)***. *A Markov chain is said to be irreducible if all the states communicate with each other, i.e. $\forall x, y \in \mathbb{X} : x \to y$. Given a distribution ν on $\mathbb{X}$, a Markov chain is ν-irreducible if for every state with positive probability under ν communicates with every other such state:*

$$\forall x, y \in supp(\nu) : x \to y$$

where $supp(\nu) = \{x \in \mathbb{X} : \nu(x) > 0\}$. It is said to be strongly irreducible if any state can be reached from any point in the space in a single step and strongly ν-irreducible if all states in $supp(\nu)$ may be reached in a single step.

This notion is important for the study of Markov chain Monte Carlo methods as a chain with this property can somehow explore the entire space rather than being restricted to a subset of it.

1.3.2 Properties of states

Another important notion is the notion of periodicity.

Definition 4 ***(Period)*** *A state x in a discrete state space Markov chain has period $d(x)$ defined as:*

$$d(x) = \gcd\{s \geq 1 : K_{xx}^s > 0\}$$

where gcd denotes the greatest common denominator. A chain possessing such a state is said to have a cycle of length $d(x)$.

Proposition 1 *All states which communicate have the same period and hence, in an irreducible Markov chain, all states have the same period.*

Proof. Assume that $x \leftrightarrow y$ so there exist paths of lengths r, s and t, respectively from $x \to y$, $y \to x$ and $y \to y$, respectively. Hence there exist paths of length $r + s$ and $r + s + t$ from x to x, hence $d(x)$ must divide $r + s$ and $r + s + t$ and consequently their difference, t. As this holds for any t corresponding to a path from $y \to y$ then $d(x)$ is a divisor of the length of any path from $y \to y$: as $d(y)$ is the gcd of all such paths by definition, it follows that $d(x) \leq d(y)$. By symmetry, we also have that $d(y) \leq d(x)$, so $d(x) = d(y)$ and the proof is complete. ■

For irreducible Markov chains, the term *periodic* corresponds to chains whose states have some common period greater than 1 whereas chains whose period is 1 are termed *aperiodic*.

We now introduce an additional random quantity which corresponds to the number of times a state is visited if a Markov chain is allowed to run for infinite time:

$$\eta_x := \sum_{k=1}^{\infty} \mathbb{I}_x(X_k).$$

We will also adopt the standard convention that, given any function ϕ, $\mathbb{E}_x(\phi(X_1, X_2, ...))$ is the expectation of ϕ under the law of the Markov chain initialised with $X_1 = x$. Similarly, if μ is some distribution over $\mathbb{X}$, then $\mathbb{E}_\mu(\phi)$ is the expectation of ϕ under the law of the process initialised with $X_1 \sim \mu$.

Definition 5 ***(Transience and Recurrence).*** *In the context of discrete state space Markov chains, a state x is termed transient if:*

$$\mathbb{E}_x(\eta_x) < \infty$$

whilst, if we have that,

$$\mathbb{E}_x(\eta_x) = \infty$$

then that state is termed recurrent.

For irreducible Markov chains, transience and recurrence are properties of the chain itself, rather than its individual states: if any state is transient (or recurrent) then all states have that property. Indeed, for an irreducible Markov chain either all states are recurrent or all are transient.

1.3.3 Equilibrium

We will be particularly interested in this course in Markov kernels admitting an invariant distribution.

Definition 6 ***(Invariant Distribution).*** *A distribution π is said to be invariant or stationary for a Markov kernel, K, if $\pi K = \pi$.*

The invariant distribution π of a Markov kernel K is simply the left eigenvector with unit eigenvalue. If there exists t such that $X_t \sim \pi$ where π is a stationary distribution, then $X_{t+s} \sim \pi K^s = \pi$ for all $s \in \mathbb{N}$. A Markov chain is said to be in its stationary regime once this has occurred. Note that this tells us nothing about the correlation between the states or their joint distribution.

Definition 7 ***(Reversibility).*** *A stationary stochastic process is said to be reversible if the statistics of the time-reversed version of the process match those of the process in the forward distribution, so that reversing time makes no discernible difference to the sequence of distributions which are obtained, that is the distribution of any collection of future states given any past history must match the conditional distribution of the past conditional upon the future being the reversal of that history.*

Numerous Markov chains we will study later on are reversible. Reversibility is typically verified by checking the detailed balance condition discussed below. If this condition holds for a distribution, then it also tells us that this distribution is the stationary distribution of the chain.

Proposition 2 ***(Detailed Balance).*** *If a Markov kernel satisfies the so-called detailed balance condition for some distribution π,*

$$\forall x, y \in \mathbb{X} : \ \pi_x K_{xy} = \pi_y K_{yx} \tag{5}$$

then

1. *π is the invariant distribution of the chain.*
2. *The chain is reversible with respect to π.*

Proof. To demonstrate that K is μ-invariant, we sum both sides of the detailed balance equation (5) over x:

$$\sum_{x \in \mathbb{X}} \pi_x K_{xy} = \sum_{x \in \mathbb{X}} \pi_y K_{yx}$$
$$(\pi K)_y = \pi_y$$

and as this holds for all y then $\pi K = \pi$.

To check that the chain is reversible, we check that

$$\begin{aligned}
\mathbb{P}(X_t = x | X_{t+1} = y) &= \frac{\mathbb{P}(X_t = x, X_{t+1} = y)}{\mathbb{P}(X_{t+1} = y)} \\
&= \frac{\pi_x K_{xy}}{\pi_y} = \frac{\pi_y K_{yx}}{\pi_y} \text{ (detailed balance)} \\
&= K_{yx} = \mathbb{P}(X_t = x | X_{t-1} = y).
\end{aligned}$$

In the case of a Markov chain it is clear that if the transitions are time-reversible then the process must be time reversible. ■

2 General State Space Markov Chains

2.1 Basic Concepts

The study of general state space Markov chains is far beyond the scope of this course. In this section, we will just explain how some of the concepts introduced for discrete state spaces can be extended to continuous domains via the use of probability densities. We will not provide proofs of results but will provide a list of references for the interested reader.

When dealing with continuous state spaces, the principle complication stems from the fact that the probability of any random variable distributed according to a non-degenerate density taking any particular value is zero. The Markov property (2) extended to a continuous state space states that for any measurable set $A \subset \mathbb{X}$:

$$\mathbb{P}(\, X_t \in A |\, X_1 = x_1, X_2 = x_2, ..., X_{t-1} = x_{t-1}) = \mathbb{P}(\, X_t \in A |\, X_{t-1} = x_{t-1}).$$

In the case which we are interested in, it is often convenient to describe the distribution of a random variable over $\mathbb{X}$ in terms of some probability density, $\mu : \mathbb{X} \to \mathbb{R}$ which has the property that, if $X \sim \mu$, then we have for any measurable set A that:

$$\mathbb{P}\left(X \in A\right) = \int_A \mu\left(x\right) dx.$$

We will restrict ourselves to the homogeneous case here but the extension to inhomogeneous Markov chains is straightforward. For homogeneous chains, we may describe the conditional probabilities of interest as a kernel function $K : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ which has the property that for all measurable sets $A \subset \mathbb{X}$ and all $x \in \mathbb{X}$:

$$\mathbb{P}(\, X_t \in A |\, X_{t-1} = x) = \int_A K\left(x, y\right) dy := K\left(x, A\right),$$

that is conditional on $X_{t-1} = x$, X_t is a random variable which admits a pdf $K\left(x, \cdot\right)$.

Hence it follows that, for any collection of measurable sets $A_1, A_2, ...$ the following holds:

$$\mathbb{P}(X_1 \in A_1, X_2 \in A_1, ..., X_t \in A_t) = \int_{A_1 \times \cdots \times A_t} \mu\left(x_1\right) \prod_{k=2}^{t} K\left(x_{k-1}, x_k\right) dx_1 \cdots dx_{k-1}.$$

We can also define the m-step ahead conditional distributions,

$$\mathbb{P}(\, X_{t+m} \in A |\, X_t = x_t) = \int_{\mathbb{X}^{m-1} \times A} \prod_{k=t+1}^{t+m} K\left(x_{k-1}, x_k\right) dx_{t+1} \cdots dx_{t+m}$$

and it is useful to define an m-step ahead transition kernel in the same manner as it is in the discrete case, here matrix multiplication is replaced by a convolution operation but the intuition remains the same; i.e. we can rewrite the expression above as

$$\mathbb{P}(\, X_{t+m} \in A |\, X_t = x_t) = \int_A K^m\left(x_t, x_{t+m}\right) dx_{t+m} := K^m\left(x_t, A\right),$$

where

$$K^m\left(x_t, x_{t+m}\right) := \int_{\mathbb{X}^{m-1}} \prod_{k=t+1}^{t+m} K\left(x_{k-1}, x_k\right) dx_{t+1} \cdots dx_{t+m-1}.$$

If we denote by μ_t the density of the marginal distribution of X_t, we obtain

$$\mu_{t+m}\left(y\right) = \int_{\mathbb{X}} \mu_t\left(x\right) K^m\left(x, y\right) dx$$

and

$$\mu_{t+m}\left(A\right) := \mathbb{P}(X_{t+m} \in A) = \int_A \left(\int_{\mathbb{X}} \mu_t\left(x\right) K^m\left(x, y\right) dx \right) dy.$$

Example 1 *Consider the following autoregressive (AR) model*

$$X_t = \rho X_{t-1} + V_t \tag{6}$$

where $V_t \overset{i.i.d.}{\sim} p_V(\cdot)$. *This defines a Markov process such that*

$$K(x,y) = p_V(y-\rho x).$$

In particular for $p_V(v) = \mathcal{N}\left(v;0,\tau^2\right)$, *we have*

$$K(x,y) = \frac{1}{\sqrt{2\pi\tau^2}}\exp\left(-\frac{1}{2\tau^2}(y-\rho x)^2\right).$$

We also have

$$X_{t+2} = \rho(\rho X_t + V_{t+1}) + V_{t+2} = \rho^2 X_t + \rho V_{t+1} + V_{t+2}$$

and similarly

$$X_{t+m} = \rho^m X_t + \sum_{k=1}^{m} \rho^{m-k} V_{t+k}.$$

So when $p_V(v) = \mathcal{N}\left(v;0,\tau^2\right)$, *we have*

$$K^m(x,y) = \frac{1}{\sqrt{2\pi\tau_m^2}}\exp\left(-\frac{1}{2}\frac{(y-\rho^m x)^2}{\tau_m^2}\right) \tag{7}$$

where

$$\tau_m^2 = \tau^2\sum_{k=1}^{m}\left(\rho^2\right)^{m-k} = \tau^2\frac{1-\rho^{2m}}{1-\rho^2}. \tag{8}$$

Example 2 *Consider the following model. At time* t, *with probability* $\alpha(X_{t-1})$ *set*

$$X_t = \rho X_{t-1} + V_t$$

where $V_t \overset{i.i.d.}{\sim} p_V(\cdot)$ *and otherwise set* $X_t := X_{t-1}$. *In this case* $(X_t)_{t\geq 1}$ *is still a Markov process but*

$$K(x,y) = \alpha(x)\ p_V(y-\rho x) + (1-\alpha(x))\ \delta_x(y)$$

where $\delta_x(y)$ *is the Dirac mass located at* x. *In this scenario, the transition kernel does not admit a density with respect to the Lebesgue measure as it has a singular component. A proper - measure theoretic - way to write the transition kernel is*

$$K(x,dy) = \alpha(x)\ p_V(y-\rho x)\,dy + (1-\alpha(x))\ \delta_x(dy)$$

and then we have

$$\mathbb{P}(X_t \in A | X_{t-1} = x) = \int_A K(x,dy).$$

We simplify the notation by essentially using the (abusive) convention that $\delta_x(dy) = \delta_x(y)\,dy$ *so that* $K(x,dy) = K(x,y)\,dy$. *In this case, we also use the notation*

$$K(x,\{x\}) = \mathbb{P}(X_t = x | X_{t-1} = x).$$

2.2 Important Properties

In this section we will introduce definitions and properties which fulfill the same role in the context of continuous state spaces as those introduced earlier for discrete state spaces. In particular, we are interested in irreducibility: we want some way of determining what class of states are reachable from one another and hence what part of $\mathbb{X}$ might be explored, with positive probability, starting from a point within such a class.

Definition 8 ***(Irreducibility).*** *Given a distribution,* μ, *over* $\mathbb{X}$, *a Markov chain is said to be* μ*-irreducible if for all points* $x \in \mathbb{X}$ *and all measurable sets* A *such that* $\mu(A) > 0$ *there exists some* t *such that:*

$$K^t(x,A) > 0.$$

If this condition holds with $t = 1$, *then the chain is said to be strongly* μ*-irreducible.*

This definition tells us whether a Markov chain is likely to be satisfactory if we are interested in approximating some property of a distribution μ using samples from this chain: if it is not μ-irreducible then there are some points in the space from which we cannot reach all of the support of μ, and this is likely to be a problem. In the sequel we will be interested more or less exclusively in irreducible Markov chains with respect to some distribution of interest.

It is necessary to be a bit subtle when extending some of the concepts introduced in the case of discrete space space to continuous state-space. In particular, it is necessary to introduce the concept of small sets; these act as a replacement for the individual states of a discrete space Markov chain. A first attempt is to consider the sets which have the property that the distribution taken by the Markov chain at time $t+1$ is the same if it starts at any point in this set – so the conditional distribution function is constant over this set.

Definition 9 ***(Atoms).*** *A Markov chain with transition kernel K is said to have an atom, $\alpha \subset \mathbb{X}$, if there is some probability distribution, ν, such that:*

$$\forall x \in \alpha, A \subset \mathbb{X} : \quad \int_A K(x,y)\, dy = \int_A \nu(y)\, dy$$

If the Markov chain in question is ν-irreducible, then α is termed an accessible atom.

The concept of atoms allow us to introduce some structure similar to that seen in discrete chains; i.e it provides us with a set of positive probability which, if the chain ever enters it, we know the distribution of the subsequent state. However most interesting continuous state space Markov chains do not possess atoms. The condition that the distribution in the next state is precisely the same is somehow too strong. An alternative approach consists in requiring that the conditional distribution has a common component. This is the intuition behind a much more useful concept which underlies much of the analysis of general state space Markov chains.

Definition 10 ***(Small Sets).*** *A set, $C \subset \mathbb{X}$, is termed small for a given Markov chain (or, when one is being precise, (ν, m, ϵ)-small) if there exists some positive integer m, some $\epsilon > 0$ and some probability distribution, ν, such that:*

$$\forall x \in \alpha, A \subset \mathbb{X} : \quad \int_A K^m(x,y)\, dy \geq \epsilon \int_A \nu(y)\, dy$$

This tells us that the distribution m steps after the chain enters the small set has a component of size at least ϵ of the distribution ν, wherever it was within that set. In this sense, small sets are not "too big": there is potentially some commonality of all paths emerging from them. Although we have not proved that such sets exist for any particular class of Markov chains it is, in fact, the case that they do for many interesting Markov chains and their existence allows for a number of sophisticated analytic techniques to be applied.

Definition 11 ***(Periodic and Aperiodic).*** *A μ-irreducible Markov chain of transition kernel K is periodic if there exists some partition of the state space, $\mathbb{X}_1, ..., \mathbb{X}_d$ for $d \geq 2$, i.e. $\forall i \neq j : \mathbb{X}_i \cap \mathbb{X}_j = \varnothing$ and $\cup_{i=1}^{d} \mathbb{X}_i = \mathbb{X}$, with the properties that:*

$$\forall i, j, t, s : \ \mathbb{P}(X_{t+s} \in \mathbb{X}_j | X_t \in \mathbb{X}_i) = \begin{cases} 1 & j = (i+s) \ mod \ d \\ 0 & otherwise. \end{cases} .$$

Otherwise the kernel is aperiodic.

What this actually tells us is that a Markov chain with a period of d is such that the chain moves with probability 1 from set $\mathbb{X}_1$ to $\mathbb{X}_2$, $\mathbb{X}_2$ to $\mathbb{X}_3$, ...,$\mathbb{X}_{d-1}$ to $\mathbb{X}_d$ and $\mathbb{X}_d$ to $\mathbb{X}_1$ (assuming that $d \geq 2$, of course). Hence the chain will visit a particular element of the partition with a period of d.

We also require some way of characterising how often a continuous state space Markov chain visits any particular region of the state space in order to obtain concepts analogous to those of transience and recurrence in the discrete setting. In order to do this we define a collection of random variables η_A for any measurable set A of $\mathbb{X}$, which correspond to the number of times the set A is visited, i.e.

$$\eta_A := \sum_{k=1}^{\infty} \mathbb{I}_A(X_k)$$

and, once again we use $\mathbb{E}_x$ to denote the expectation under the law of the Markov chain with initial state x. We note that if a chain is not μ-irreducible for some distribution μ, then there is no guarantee that it is either transient or recurrent, however, the following definitions do hold:

Definition 12 ***(Transience and Recurrence)****. We begin by defining uniform transience and recurrence for sets $A \subset \mathbb{X}$ for μ-irreducible general state space Markov chains. Such a set is recurrent if:*

$$\forall x \in A : \mathbb{E}_x(\eta_A) = \infty.$$

A set is uniformly transient if there exists some $M < \infty$ such that:

$$\forall x \in A : \mathbb{E}_x(\eta_A) \leq M.$$

The weaker concept of transience of a set may then be introduced. A set, $A \subset \mathbb{X}$, is transient if it may be expressed as a countable union of uniformly transient sets, i.e.:

$$\exists \{B_i \subset \mathbb{X}\}_{i=1}^{\infty} : A \subset \cup_{i=1}^{\infty} B_i$$
$$\forall i \in \mathbb{N} : \forall x \in B_i : \mathbb{E}_x(\eta_{B_i}) \leq M_i < \infty.$$

A general state space Markov chain is recurrent if the following two conditions are satisfied:
- *The chain is μ-irreducible for some distribution μ.*
- *For every measurable set $A \subset \mathbb{X}$ such that $\mu(A) = \int_A \mu(x)\,dx > 0$, $\mathbb{E}_x(\eta_A) = \infty$ for every $x \in A$.*

The chain is transient if it is μ-irreducible for some distribution μ and the entire space is transient.

As in the discrete setting, in the case of irreducible chains, transience and recurrence are properties of the chain rather than individual states: all states within the support of the irreducibility distribution are either transient or recurrent. It is useful to note that any μ-irreducible Markov chain which has stationary distribution μ is positive recurrent (Tierney, 1994). A slightly stronger form of recurrence is widely employed in the proof of many theoretical results which underlie many applications of Markov chains to statistical problems. This form of recurrence is known as Harris recurrence and may be defined as follows:

Definition 13 ***(Harris Recurrence)****. A set $A \subset \mathbb{X}$ is Harris recurrent if $\mathbb{P}_x(\eta_A = \infty) = 1$ for every $x \in \mathbb{X}$. A Markov chain is Harris recurrent if there exists some distribution μ with respect to which it is irreducible and every set A such that $\mu(A) > 0$ is Harris recurrent.*

The concepts of invariant distribution, reversibility and detailed balance are essentially unchanged from the discrete setting. It is necessary to consider integrals with respect to densities rather than sums over probability distributions, but no fundamental differences arise here.

Definition 14 ***(Invariant Distribution)****. A distribution of density π is said to be invariant or stationary for a Markov kernel, K, if*

$$\int_{\mathbb{X}} \pi(x) K(x, y)\,dx = \pi(y).$$

Proposition 3 *If a Markov kernel satisfies the so-called detailed balance condition for some distribution of density π, if for any $x, y \in \mathbb{X}$*

$$\forall x, y \in \mathbb{X} : \ \pi(x) K(x, y) = \pi(y) K(y, x)$$

then
1. *π is the invariant distribution of the chain.*
2. *The chain is reversible with respect to π.*

Example 3 *For the autoregressive (AR) Gaussian model (6), we can easily check that the detailed balance condition is satisfied for*

$$\pi(x) = \mathcal{N}\left(x; 0, \frac{\tau^2}{1-\rho^2}\right)$$

when $|\rho| < 1$. We can also easily check using (7)-(8) that in this case we have

$$\lim_{t \to \infty} K^t(x, y) = \pi(y)$$

so that the distribution of X_t is becoming independent of $X_1 = x$.

3 Selected Theoretical Results

The probabilistic study of Markov chains dates back more than fifty years and comprises an enormous literature, much of it rather technically sophisticated. We do not intend to summarise that literature here, nor to provide proofs of the results which we present here. This section serves only to motivate the material presented in the subsequent chapters.

These theorems fill the role which the law of large numbers and the central limit theorem for independent, identically distributed random variables fill in the case of simple Monte Carlo methods. They tell us, roughly speaking, that if we take the sample averages of a function at the points of a Markov chain which satisfies suitable regularity conditions and possesses the correct invariant distribution, then we have convergence of those averages to the integral of the function of interest under the invariant distribution and, furthermore, under stronger regularity conditions we can obtain a rate of convergence.

There are two levels of strength of law of large numbers which it is useful to be aware of. The first tells us that for most starting points of the chain a law of large numbers will hold. Under slightly stronger conditions (which it may be difficult to verify in practice) it is possible to show the same result holds for all starting points.

Theorem 15 ***(A Simple Ergodic Theorem)****. If K is a π-irreducible, recurrent $\mathbb{R}^d$-valued Markov kernel which admits π as a stationary distribution, then the following strong law of large numbers holds (convergence is with probability 1) for any integrable function $\phi : \mathbb{X} \to \mathbb{R}$:*

$$\lim_{t\to\infty} \frac{1}{t} \sum_{i=1}^{t} \phi(X_i) = \int_{\mathbb{X}} \phi(x)\, \pi(x)\, dx$$

for $\pi-$almost all starting value x. That is, for any x except perhaps for some set $\mathcal{N}$ such that $\int_{\mathcal{N}} \pi(x)\, dx = 0$.

An outline of the proof of this theorem is provided by (Roberts and Rosenthal, 2004, Fact 5.).

Theorem 16 ***(A Stronger Ergodic Theorem).*** *If K is a π-invariant, Harris recurrent Markov chain, then the following strong law of large numbers holds (convergence is with probability 1) for any integrable function $\phi : \mathbb{X} \to \mathbb{R}$:*

$$\lim_{t\to\infty} \frac{1}{t} \sum_{i=1}^{t} \phi(X_i) = \int_{\mathbb{X}} \phi(x)\, \pi(x)\, dx$$

A proof of this result is beyond the scope of the course. This is a particular case of (Robert and Casella, 2004, p. 241, Theorem 6.63), and a proof of the general theorem is given there. The same theorem is also presented with proof in (Meyn and Tweedie, 1993, p. 433, Theorem 17.3.2).

Note that the previous results do not ensure that $X_t \sim \pi$ as $t \to \infty$. An additional condition, namely aperiodicity, is necessary to ensure this.

Theorem 17 *Suppose the kernel K is π-irreducible, π-invariant and aperiodic. Then, we have*

$$\lim_{t\to\infty} \int_{\mathbb{X}} \left| K^t(x, y) - \pi(y) \right| dy = 0$$

for $\pi-$almost all starting value x.

Theorem 18 ***(A Central Limit Theorem)****. Under technical regularity conditions (see (Jones, 2004) 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, π-invariant Markov chain, and a function $\phi : \mathbb{X} \to \mathbb{R}$ which has at least two finite moments (depending upon the combination of regularity conditions assumed, it may be necessary to have a finite moment of order $2 + \delta$)*

$$\lim_{t\to\infty} \sqrt{t} \left[\frac{1}{t} \sum_{i=1}^{t} \phi(X_i) - \int_{\mathbb{X}} \phi(x)\, \pi(x)\, dx \right] \xrightarrow{D} \mathcal{N}\left(0, \sigma^2(\phi)\right),$$

$$\sigma^2(\phi) = \mathbb{V}\left[\phi(X_1)\right] + 2 \sum_{k=2}^{\infty} \mathbb{C}ov\left[\phi(X_1), \phi(X_k)\right]$$

where the variance and covariance in the expression above are with respect to the distribution of the Markov chain in its stationary regime.

Example 4 *For the autoregressive (AR) Gaussian model (6), we have seen that*

$$\pi(x) = \mathcal{N}\left(x; 0, \frac{\tau^2}{1-\rho^2}\right)$$

when $|\rho| < 1$. *We can also easily check that in the stationary regime*

$$\begin{aligned}\mathbb{C}ov(X_1, X_k) &= \rho \mathbb{C}ov(X_1, X_{k-1}) = \cdots = \rho^{k-1}\mathbb{V}[X_1] \\ &= \rho^{k-1}\frac{\tau^2}{1-\rho^2}.\end{aligned}$$

Hence we can easily check that the variance in the CLT for $\phi(x) = x$ *is given by*

$$\begin{aligned}\sigma^2 &= \mathbb{V}(X_1) + 2\sum_{k=2}^{\infty}\mathbb{C}ov(X_1, X_k) \\ &= \frac{\tau^2}{1-\rho^2}\left(1 + 2\sum_{k=1}^{\infty}\rho^k\right) \\ &= \frac{\tau^2}{1-\rho^2}\frac{1+\rho}{1-\rho}.\end{aligned}$$

References

[1] G.L. Jones, On the Markov chain central limit theorem, *Probability surveys*, vol. 1, pp. 299–320, 2004.

[2] S. Meyn & R. Tweedie, *Markov chains and Stochastic Stability*, Cambridge University Press, 1993.

[3] C.P. Robert & G. Casella, *Monte Carlo Statistical Methods*, Springer, 2004.

[4] G.O. Roberts & J.S. Rosenthal, General State-Space Markov chains and MCMC Algorithms, *Probability Surveys*, vol. 4, pp. 20-71, 2004.

[5] L. Tierney, Markov chains for exploring posterior distributions, *Annals of Statistics*, vol. 22, pp. 1701-1762, 1994.