Advanced Simulation Methods

Chapter 4 - Elements of Markov Chains Theory

We present a very brief and elementary introduction to the theory of Markov chains. These theoretical results are crucial to justify Markov chain Monte Carlo methods that will be presented in the following lectures. Markov chain Monte Carlo are the building blocks of advanced Monte Carlo methods, that address the problem of approximating integrals in high dimension. A rigorous treatment of this theory would require measure theoretic concepts, which are beyond the scope of this course. We will thus neglect these important issues here, and try to preserve the main ideas.

1 Discrete-time 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 some 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 characterize 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

$$\mathbb{P}\left(X_{t_1} \in A_{t_1}, X_{t_2} \in A_{t_2}, ..., X_{t_n} \in A_{t_n}\right).$$

The fact that those probabilities completely define a stochastic process is not obvious, because the stochastic process is "infinitely long", i.e. $t \in \mathbb{N}$ takes infinitely many values. The characterization by finite-dimensional distributions is called the Kolmogorov extension theorem. In other words, this theorem allows the definition of 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. Thus, 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 stochastic processes called "Markov", which are useful in the context of Monte Carlo methods. We will see that their specification only requires an "initial distribution" and a "transition probability" or "transition kernel", both of which are conceptually simple objects.

2 Discrete State Space Markov Chains

2.1 Markov property

Let us first consider 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 probability of the process taking a particular value at a particular time t. This contrasts with the case of continuous state spaces, where a random variable admitting a probability density function with respect to the Lebesgue measure has zero probability of taking any particular value. 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. To simplify the process, we can assume that the distribution of X_s given its "past" $(X_1, \dots, X_{s-1})$ depends only upon X_{s-1}; i.e. we have

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

Stochastic processes for which (2) holds are known as Markov processes, or simply as Markov chains, in the Monte Carlo literature. We will drop the "discrete-time" phrase from now on, but note that Markov processes can also be defined in continuous time. The fact that X_t depends only on X_{t-1} is often called the "Markov property".

When dealing with discrete state spaces, it is often convenient to associate each probability distribution with a row vector, with non-negative entries and summing to one. Now, given a random variable X on $\mathbb{X}$, we say that X has distribution μ for some vector μ (with non-negative entries and summing to one), and we note:

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

2.2 Homogeneous Markov Chains

Markov chains are called homogeneous when the 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 transition matrix $K(i, j) = K_{ij} = \mathbb{P}(X_t = j | X_{t-1} = i)$. K is also 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 by combining (1)-(2)-(3), the joint distribution of the chain over any finite time steps 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:

$$K^n_{ij} = \mathbb{P}(X_{t+n} = j | X_{t-1} = i).$$

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

$$K^{m+n}_{ij} = \sum_{k \in \mathbb{X}} K^m_{ik} K^n_{kj}.$$

which proves that indeed K^n is the n^{th} matrix power of K, and hence the notation is consistent with standard linear algebra. For the marginal laws of X_t, we obtain the expression

$$\mu_{t+1}(j) = \sum_{i \in \mathbb{X}} \mu_t(i) K_{ij}.$$

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

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

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

To summarize, homogeneous Markov chains can be characterized as follows. First, the distribution μ_0 of X_0, called the "initial distribution", must be specified. Then, the transition kernel K must be specified, and it characterizes 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 characterize stochastic processes, i.e. Kolmogorov extension theorem.

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 analyze, hence we will restrict ourselves to homogeneous chains henceforth.

2.3 Important Properties

2.3.1 Irreducibility

We review here some of the main concept/properties associated to a Markov chain $(X_t)_{t\in\mathbb{N}}$. We first consider how states communicate to each other under a given Markov chain transition kernel.

Definition 2.1. ***(Accessibility)***. *A state y is accessible from a state x, written "$x \to y$", if*

$$\inf\{t : \mathbb{P}(X_t = y | X_1 = x) > 0\} < \infty.$$

Note that this can be rewritten equivalently as $\inf\left\{t : K^t_{xy} > 0\right\} < \infty$. 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, according to the Markov kernel K.

Definition 2.2. ***(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 the study of the "communication structure" of a Markov chain, i.e. from which points it is possible to travel 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 2.3. ***(Irreducibility)***. *A Markov chain is said to be irreducible if all the states communicate, i.e. $\forall x, y \in \mathbb{X} : x \to y$. Given a probability distribution ν on $\mathbb{X}$, a Markov chain is said to be ν-irreducible if every state with positive probability under ν communicates with every other such state:*

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

where $supp(\nu) = \{x \in \mathbb{X} : \nu(x) > 0\}$. A Markov chain is said to be strongly irreducible if any state can be reached from any other state, in only one step of the Markov chain. A Markov chain is said to be strongly ν-irreducible if all states in $supp(\nu)$ may be reached in a single step from any other state in $supp(\nu)$.

This notion is important for the study of Markov chain Monte Carlo methods: indeed a Markov chain that is ν-irreducible can explore the entire support of ν, rather than being restricted to a subset of it. Thus, when we will introduce Markov chains to explore a particular distribution of interest π, we will carefully check whether the chains are π-irreducible.

2.3.2 Properties of states

Another important notion is the notion of periodicity.

Definition 2.4. ***(Periodicity)*** *For a Markov chain with kernel K, a state x has period $d(x)$ defined as:*

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

where "gcd" denotes the greatest common denominator. If a chain induces a state x of period $d(x) > 1$, it is said to have a cycle of length $d(x)$.

Proposition 2.1. *All states that communicate have the same period, therefore all states have the same period if the Markov chain is irreducible.*

Proof. Assume that x and y communicate: $x \to y$ and $y \to x$. There exist paths of lengths r, s and t, respectively from x to y, from y to x and from y to y. Hence there exist paths of length $r + s$ and $r + s + t$ from x to x, hence $d(x)$ must divide $r + s$ and $r + s + t$. Thus $d(x)$ divides their difference, that is 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 $d(y) \leq d(x)$. Therefore $d(x) = d(y)$. ■

The proposition shows that for irreducible Markov chains, we can talk about the period of the chain, defined as the period of any of its state. Then, the term *periodic* corresponds to chains with a period greater than 1, while chains with a period equal to 1 are termed *aperiodic*.

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

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

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

Definition 2.5. ***(Transience and Recurrence).*** *For a Markov chain, a state x is termed transient if:*

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

Otherwise the state is called recurrent and

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

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.

Example 2.1. *Consider the random walk on $\mathbb{Z}^k$, for an integer $k \in \mathbb{N}$. It is shown in [6], Section 1.6, that the random walk is recurrent on $\mathbb{Z}$ and $\mathbb{Z}^2$ but is transient on $\mathbb{Z}^k$ for $k \geq 3$.*

2.3.3 Equilibrium

For Markov chain Monte Carlo methods, we will be particularly interested in Markov kernels admitting an invariant distribution.

Definition 2.6. ***(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 thus 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 $s \in \mathbb{N}$, i.e. the Markov chain then keeps the same marginal distribution π forever. A Markov chain is said to be in its stationary regime once this has occurred.

The very remarkable property of aperiodic, irreducible Markov chains with an invariant distribution π, is that the marginal distribution of the chain X_t converges to the invariant distribution π. This is true for any starting state x or starting distribution μ. Hence the invariant distribution is also called the equilibrium distribution, or limiting distribution. The main result is given in the following theorem, proven in [6], Section 1.8 (Theorem 1.8.3).

Theorem 2.1. ***(Convergence to equilibrium).*** *Consider a Markov chain $\left(X_t \right)_{t \in \mathbb{N}}$ on a discrete space $\mathbb{X}$ that is aperiodic, irreducible and with an invariant distribution π. Let μ be any initial distribution on $\mathbb{X}$. Then*

$$\forall x \in \mathbb{X} \quad \mathbb{P}_\mu \left(X_t = x \right) \xrightarrow[t \to \infty]{} \pi(x).$$

This theorem is remarkable in many respects. Note that the convergence occurs for any initial distribution μ; in other words, the initial distribution is progressively "forgotten" by the chain. Note also that the theorem implies that the invariant distribution π is unique. For the construction of Markov chain Monte Carlo methods, we will typically have a specific distribution π in mind, and the goal is going to construct a Markov chain (and thus its transition kernel), such that its invariant distribution is precisely π. The "detailed balance" condition is a tool that allows the design of a transition kernel given a specific invariant distribution π.

Definition 2.7. *A Markov kernel K satisfies the detailed balance condition for a distribution π if*

$$\forall x, y \in \mathbb{X} \quad \pi(x) K_{xy} = \pi \left(y \right) K_{yx}. \tag{5}$$

We have the following straightforward proposition.

Proposition 2.2. *If a Markov kernel K satisfies the detailed balance condition for π, then π is invariant for K.*

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

$$\forall y \in \mathbb{X} \quad \sum_{x \in \mathbb{X}} \pi(x) K_{xy} = \sum_{x \in \mathbb{X}} \pi(y) K_{yx}$$
$$\text{hence} \quad \forall y \in \mathbb{X} \quad (\pi K)(y) = \pi(y)$$

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

This proposition shows that if one finds a Markov kernel K satisfying the detailed balance condition for π, and if additionally K is aperiodic and irreducible, then we can apply the convergence to equilibrium theorem, for any starting value or initial distribution.

Finally, another useful property of Markov chain is reversibility, or "time-reversibility". Essentially, a reversible Markov chain is a process that behaves similarly when considered "forward in time" or "backward in time".

Definition 2.8. ***(Time-reversal Markov chain)***. *Consider an irreducible Markov chain $(X_t)_{t \in \mathbb{N}}$, with kernel K and invariant distribution π. Assume that $X_1 \sim \pi$, i.e. the chain is started at stationarity. Define $Y_t = X_{T-t}$ for some $T \geq 1$. Then $(Y_t)_{0 \leq t \leq T}$ is an irreducible Markov chain, with invariant distribution π and with initial distribution also π, and its transition kernel $\hat{K}$ is given by*

$$\forall x, y \in \mathbb{X} \quad \pi(x) K_{xy} = \pi(y) \hat{K}_{yx}.$$

The chain (Y_t) is called the time-reversal chain of (X_t).

Definition 2.9. ***(Reversible chain)***

A irreducible Markov chain $(X_t)_{t \in \mathbb{N}}$, with kernel K and invariant distribution π, is said to be reversible (with respect to π) if its transition kernel K is equal to the transition kernel $\hat{K}$ of its time-reversal chain, in other words, at stationarity we have

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

Numerous Markov chains that we will study later on are reversible. Reversibility is typically verified by checking the detailed balance condition as discussed above.

Proposition 2.3. ***(Detailed balance implies reversibility).*** *If a Markov chain has a transition kernel K satisfying the detailed balance condition for some distribution π (as in Eq. (5)), then the chain is reversible with respect to π.*

Proof. We check that

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

■

Beyond convergence to equilibrium, there are many convergence results for Markov chains in the form of law of large numbers and central limit theorems. Those results will be of primary interest to assess the consistency and the precision of Markov chain Monte Carlo methods. Before listing some key convergence results for Markov chains, we describe elements of the theory of Markov chains defined in continuous (or "general") state spaces (such as $\mathbb{R}$).

3 Continuous State Space Markov Chains

3.1 From discrete to continuous spaces

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 spaces via probability density functions and measures. We will not provide proofs of the results, but the standard book on this topic is "the Meyn & Tweedie", as listed in the references.

When facing with continuous state spaces, the main difficulty stems from the fact that the probability of any continuous random variable taking any particular value is zero. For example, if a random variable X follows normal distribution, then for any $x \in \mathbb{R}$, $\mathbb{P}(X = \{x\}) = 0$. Hence, we cannot for instance refer to transition probabilities $\mathbb{P}(X_t = \{y\} \mid X_{t-1} = x)$. The Markov property (2) can still be defined on a continuous state space, as follows. We say that the process $(X_t)_{t\in\mathbb{N}}$ is a Markov chain if for any measurable set $A \subset \mathbb{X}$:

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

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

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

We will only consider the homogeneous case here but the extension to non-homogeneous 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 $x \in \mathbb{X}$:

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

that is conditional on $X_{t-1} = x$, X_t is a random variable which admits a probability density function $y \mapsto K(x, y)$.

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

$$\mathbb{P}(X_1 \in A_1, X_2 \in A_2, ..., X_t \in A_t) = \int_{A_1\times\cdots\times A_t} \mu(x_1) \prod_{k=2}^{t} K(x_{k-1}, x_k)\, dx_1 \cdots dx_t.$$

We can also define the m-step conditional distributions,

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

and it is useful to define an m-step transition kernel in the same manner as 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

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

where

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

Denoting by μ_t the density of the marginal distribution of X_t, we obtain

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

and, in terms of sets,

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

Example 3.1. *Consider the following auto-regressive (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 3.2. *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, because it has a singular component. A proper measure theoretic way of writing 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*

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

3.2 Important Properties

We can introduce definitions and properties which fulfill the same role in 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 3.1. ***(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.

In practice, we will deal with π-irreducible Markov chains, where π is the "target" distribution of interest. Periodicity for continuous space Markov chains can be introduced as follows.

Definition 3.2. ***(Periodicity).*** *A μ-irreducible Markov chain with transition kernel K is said to be 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 property*

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

Otherwise the Markov chain is aperiodic.

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

We need some way of characterizing how often a continuous space Markov chain visits any particular region of the state space, in order to define concepts that are analogous to transience and recurrence in the discrete space setting. Hence we introduce a collection of random variables η_A for any measurable set A of $\mathbb{X}$, which represent to the number of times that the set A is visited by the chain, i.e.

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

Once again we use $\mathbb{E}_x$ to denote the expectation under the law of the Markov chain with initial state x.

Definition 3.3. ***(Transience and Recurrence).*** *For a μ-irreducible Markov chain, a set $A \subset \mathbb{X}$ is recurrent if*

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

A set $A \subset \mathbb{X}$ is uniformly transient if there exists some $M < \infty$ such that:

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

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

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

A 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. We next introduce the concept of invariant distribution.

Definition 3.4. ***(Invariant Distribution).*** *A probability distribution with density π is said to be invariant or stationary for a Markov kernel K, if*

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

A Markov kernel that admits an invariant probability distribution and that is μ-irreducible for some μ is said to be positive recurrent. Then, it can be shown that the invariant probability distribution is unique. A slightly stronger form of recurrence is widely employed in the proof of many theoretical results which underlie Markov chain Monte Carlo. This form of recurrence is known as Harris recurrence and may be defined as follows.

Definition 3.5. ***(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 it is μ-irreducible and if every set A such that $\mu(A) > 0$ is Harris recurrent.*

The concepts of 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. For instance we can introduce reversibility as follows.

Definition 3.6. ***(Reversibility)****. A Markov kernel K is reversible with respect to π if*

$$\forall f : \mathcal{B}\left(\mathbb{X}^2 \to \mathbb{R}\right) \quad \int\int f(x,y)\pi(dx)K(x,dy) = \int\int f(y,x)\pi(dx)K(x,y),$$

where $\mathcal{B}\left(\mathbb{X}^2 \to \mathbb{R}\right)$ refers to the bounded functions measurable functions f from $\mathbb{X}^2$ to $\mathbb{R}$.

This definition, which makes apparent that the functions $(x,y) \mapsto f(x,y)$ and $(x,y) \mapsto f(y,x)$ have the same expectation with respect to two consecutive steps of the Markov chain at stationarity, is another way of expressing that the Markov chain (at stationarity) has the same law "forward" and "backward".

The detailed balance condition proves to be useful, as in the discrete setting, to design Markov kernels with a specific invariant distribution π in mind. The following result illustrates it.

Proposition 3.1. *A Markov kernel satisfies the so-called detailed balance condition for some distribution of density π, if*

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

The the following holds:

- *the distribution π is invariant for the kernel K,*
- *the chain is reversible with respect to π.*

Example 3.3. *For the auto-regressive (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 $X_1 = x$.

4 Selected Convergence Results

The study of Markov chains dates back more than fifty years and comprises an enormous literature. This section serves only to motivate the material presented in the subsequent chapters; that is, in a nutshell, Markov chains can indeed be used to approximate integrals.

These theorems are essentially laws of large numbers and central limit theorems, but for Markov chains instead of independent, identically distributed random variables. They tell us that if we take a sample average of a function evaluated at the realizations of a Markov chain, then the averages converge to the integral of the function of interest with respect to the stationary distribution of the Markov chain; at least under some "regularity conditions", on the function and on the invariant distribution. Under some stronger conditions, we can even obtain a rate of convergence.

Let us start with the law of large number, which comes in two flavors. The first tells us that, for most starting points of the chain, a law of large numbers holds. Under slightly stronger conditions which may be difficult to prove in practice, the same result holds for all starting points.

Theorem 4.1. ***(A Simple Ergodic Theorem)****. If K is a π-irreducible $\mathbb{X}$-valued Markov kernel with stationary distribution π, then almost surely (i.e. with probability one), 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 π-almost all starting values x (i.e. for any x except 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 4.2. ***(A Stronger Ergodic Theorem).*** *If K is a π-invariant, Harris recurrent Markov chain with stationary distribution π, then almost surely, 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 all starting values x.

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, aperiodicity, is necessary to ensure this.

Theorem 4.3. *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.

Laws of large numbers for stochastic processes are also called "ergodic theorems", the most famous one being Birkhoff ergodic theorem, which generalizes the above theorems. Before stating a central limit theorem, we mention different definitions of ergodicity.

Definition 4.1. ***(Geometric Ergodicity)****. A π-invariant, Harris recurrent Markov chain with stationary distribution π is geometrically ergodic if there exists a real $\rho < 1$ and a non-negative function $M : \mathbb{X} \to \mathbb{R}^+$, such that for all measurable set A,*

$$|K^n(x, A) - \pi(A)| \leq M(x)\rho^n.$$

Geometric ergodicity is thus about convergence to equilibrium, with a specific geometric rate. The bound still depends on the starting point x. Uniform ergodicity is a stronger property.

Definition 4.2. ***(Uniform Ergodicity)****. A π-invariant, Harris recurrent Markov chain with stationary distribution π is geometrically ergodic if there exists a real $\rho < 1$ and a real $M > 0$, such that for all measurable set A,*

$$|K^n(x, A) - \pi(A)| \leq M\rho^n.$$

Under regularity conditions, given e.g. in Jones, 2004, 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 $2 + \delta$). It can be proved that every reversible, geometrically ergodic, stationary Markov chain satisfies a central limit theorem, which is why geometric ergodicity was introduced above as an intermediate concept.

Theorem 4.4. ***(A Central Limit Theorem)****. For a Harris recurrent, π-invariant Markov chain, and a function $\phi : \mathbb{X} \to \mathbb{R}$ satisfying enough regularity conditions,*

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

where

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

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

Example 4.1. *For the auto-regressive (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.

[6] J.R. Norris, Markov chains. Cambridge university press, 1998.