1 General State Space MC Theory

1.1 Introduction

Markov chain theory is usually introduced in a course on stochastic processes under the restriction that the state space is countable. In order to apply Markov chain theory to most Bayesian problems (to sample from the posterior) we require a generalization of this theory to general state spaces. Usually we are interested in Euclidean space (e.g. $\mathbb{R}$ or $\mathbb{R}^d$ or a relevant subspace), however not much more effort is required to work in non-topological spaces.

Markov chains are constructed from a transition kernel P. A transition kernel is a conditional probability measure such that $\Phi_n \sim P(\cdot \mid \Phi_{n-1})$. The chains in MCMC settings have a strong stability property. A stationary probability distribution exists by construction of the chain. That is a distribution π exists such that for some N, if $\Phi_n \sim \pi \implies \Phi_{n+1} \sim \pi, \forall\, n > N$, if P allows moves over the entire state space. This freedom is called irreducibility. Irreducibility ensures that most MCMC chains are recurrent (i.e. the average number of visits to an arbitrary set B is infinite) or Harris recurrent (the probability of an infinite number of returns to B is 1). Harris recurrence ensures that the chain has the same limiting behavior for every starting value instead of almost every starting value. This is quite important as in practice we start chains from an arbitrary point x_0: a set of measure zero (under a continuous dominating measure).

The stationary distribution, π, is also a limiting distribution. That is, $\pi(\Phi_n) \to \pi$ under the total variation norm, independently of the initial state, x_0. Stronger forms of convergence also appear in MCMC theory, such as geometric and uniform convergence. A consequence of this convergence is that

$$\lim_{N\uparrow\infty} \left(\frac{1}{N} \sum_{n=1}^{N} h(\Phi_n) \right) = \mathbb{E}_\pi[h(\Phi)], \quad a.s.$$

If the chain is time reversible (i.e. symmetric w.r.t. time), then a Central Limit Theorem holds. We begin with some definitions in the theory of stochastic processes.

1.2 Stochastic Processes

Let $\mathcal{X}$ denote a general space and $\mathcal{B}(\mathcal{X})$ denote a σ-algebra of subsets of $\mathcal{X}$. Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ be a measurable space.

Definition 1 (Random Variable) *A mapping $X : \Omega \to \mathcal{X}$ is called a random variable if*

$$X^{-1}(A) := \{\omega : X(\omega) \in A\} \in \mathcal{F}, \quad \forall A \in \mathcal{B}(\mathcal{X}).$$

Note that this mapping induces a probability measure on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. Specifically

$$\Pr(X \in A) := P(X^{-1}(A) \in \mathcal{F}).$$

We will write $P(X \in A) \equiv \Pr(X \in A)$ although technically the probability P on $(\Omega, \mathcal{F})$ and the probability measure induced by the mapping are different.

Definition 2 (Stochastic Process) *A stochastic process with state space $\mathcal{X}$ is a collection of random variables indexed by a set T, $\mathbf{\Phi} = \{\Phi(t) : t \in T\}$. That is for $A \in \mathcal{B}(\mathcal{X})$, $P(\Phi(t) \in A) := P(\{\omega : \Phi(\omega, t) \in A\} \in \mathcal{F}) = P(\Phi^{-1}(A, t) \in \mathcal{F})$.*

Definition 3 (Discrete Time Stochastic process) *A discrete time stochastic process is a stochastic process where T is the non-negative integers $\mathbb{N}_+$.*

We will only be concerned with discrete time stochastic processes and will denote $\Phi(t) \equiv \Phi_t$.

We can also think of the whole of a discrete time stochastic process $\mathbf{\Phi}$ as an entity in its own right, called sample paths or realizations of the process, lying in the product space $\mathcal{X}^\infty = \prod_{i=0}^{\infty} \mathcal{X}_i$, where $\mathcal{X}_i \equiv \mathcal{X}$, each equipped with $\mathcal{B}(\mathcal{X})$.

Definition 4 (Countable State Space) *The state space $\mathcal{X}$ is called countable, if $\mathcal{X}$ is discrete with a finite or countable number of elements. $\mathcal{B}(\mathcal{X})$ is the σ-algebra generated by all subsets of $\mathcal{X}$.*

Definition 5 (Topological State Space) *The state space $\mathcal{X}$ is called topological if it equipped with a locally compact, separable, metrizable topology, then $\mathcal{B}(\mathcal{X})$ is the Borel σ-algebra (the smallest σ-algebra generated by the open sets).*

Definition 6 (General State Space) *The state space $\mathcal{X}$ is call general if it is equipped with a countably generated σ-algebra $\mathcal{B}(\mathcal{X})$.*

For the most part, we will be concerned mainly with general state spaces.

1.3 Probability Transition Kernels

A probability transition kernel plays the same role in general state space Markov chains as the probability transition matrix plays in countable state spaces. Let $\mathbb{R}_+$ denote the non-negative real numbers.

Definition 7 (Probability Transition Kernel) *A probability transition kernel is a function, P such that*

1. $\forall\, x \in \mathcal{X}$, $P(x, \cdot)$ *is a probability measure on* $\mathcal{B}(\mathcal{X})$*; in this case* $P(x, \cdot) : \mathcal{B}(\mathcal{X}) \to [0, 1]$
2. $\forall\, A \in \mathcal{B}(\mathcal{X})$, $P(\cdot, A)$ *is a non-negative measurable function on* $\mathcal{X}$*.*

When $\mathcal{X}$ is discrete, the transition kernel reduces to a transition matrix with elements

$$P(x, y) = \Pr(\Phi_n = y \mid \Phi_{n-1} = x), \quad x, y \in \mathcal{X}.$$

As a consequence of the Radon-Nikodym theorem, when $P(x, \cdot)$ is absolutely continuous w.r.t. some measure ν, there exists a function f, such that

$$P(x, A) = \Pr(\Phi_{i+1} \in A \mid \Phi_i = x) = \int_A f d\nu.$$

The function f is called the density of P w.r.t. ν and is commonly written as $f = dP/d\nu$.

Definition 8 (Markov Chain) *Given a probability transition kernel P, $\mathbf{\Phi} = \{\Phi_0, \Phi_1, \dots\}$ is a Markov chain if*

$$\Pr(\Phi_{i+1} \in A \mid \Phi_0, \dots, \Phi_i; \Phi_i = x) = \Pr(\Phi_{i+1} \in A \mid \Phi_i = x) = P(x, A).$$

If

$$\Pr(\Phi_{i+1} \in A_1, \dots, \Phi_{i+k} \in A_k \mid \Phi_i = x) = \Pr(\Phi_1 \in A_1, \dots, \Phi_k \in A_k \mid \Phi_0 = x)$$

for all i and k, then the Markov chain is said to be homogeneous.

1.3.1 The n-step probability transition kernel

Let $P^0(x, A) = \delta_x(A)$, the Dirac measure defined by $\delta_x(A) = 1$ if $x \in A$ and $\delta_x(A) = 0$ if $x \in A^c$. Then for $n \geq 1$ define

$$P^n(x, A) = \int_{\mathcal{X}} P(x, dy) P^{n-1}(y, A), \quad x \in \mathcal{X},\, A \in \mathcal{B}(\mathcal{X}).$$

P^n is called then n-step probability transition kernel.

Theorem 1 (Chapman-Kolmogorov Equations) *For any integer $m \in [0, n]$,*

$$P^n(x, A) = \int_{\mathcal{X}} P^m(x, dy) P^{n-m}(y, A), \quad x \in \mathcal{X},\ A \in \mathcal{B}(\mathcal{X}).$$

In words, the Chapman-Kolmogorov Equations state, as $\mathbf{\Phi}$ moves from x into A in n steps it must take some value $y \in \mathcal{X}$ at some intermediate time m. Since Φ is a Markov chain, it forgets its past at m and moves $n - m$ steps with the appropriate law starting at y.

We will often used the following definition:

$$P_x(\Phi_n \in A) := P^n(x, A)$$

and so the Chapman-Kolmogorov Equations can be written

$$P_x(\Phi_n \in A) = \int_{\mathcal{X}} P_x(\Phi_m \in dy) P_y(\Phi_{n-m} \in A).$$

Also, we will let μ denote the initial distribution of the chain so that

$$\begin{aligned}
&P_\mu(\Phi_0 \in A_0, \Phi_1 \in A_1, \ldots, \Phi_n \in A_n) \\
&= \int_{y_0 \in A_0} \Pr(\Phi_1 \in A_1, \ldots, \Phi_n \in A_n \mid y_0) \mu(dy_0) \\
&= \int_{y_0 \in A_0} \int_{y_1 \in A_1} \Pr(\Phi_2 \in A_2, \ldots, \Phi_n \in A_n \mid y_1) \mu(dy_0) P(y_0, dy_1) \\
&\qquad\qquad \vdots \\
&= \int_{y_0 \in A_0} \cdots \int_{y_n \in A_n} \mu(dy_0) P(y_0, dy_1) \cdots P(y_{n-1}, dy_n).
\end{aligned}$$

The m-step chain $\mathbf{\Phi}^m = \{\Phi_n^m\}$ is a sub-chain of the original chain $\mathbf{\Phi}$ with transition probabilities

$$P_x(\Phi_n^m \in A) = P^{mn}(x, A).$$

Definition 9 (Skeletons and Resolvents) *The chain $\mathbf{\Phi}^m$ with transition law $P_x(\Phi_n^m \in A) = P^{mn}(x, A)$ is called the m-skeleton of the chain Φ.*

The resolvent K_ϵ, $\epsilon \in (0, 1)$, is defined by

$$K_\epsilon(x, A) := (1 - \epsilon) \sum_{i=0}^{\infty} \epsilon^i P^i(x, A), \quad x \in \mathcal{X},\ A \in \mathcal{B}(\mathcal{X}).$$

The Markov chain with probability transition kernel K_ϵ is the called the K_ϵ-chain.

Given an initial distribution μ for the chain Φ, the K_ϵ-chain is a sub-chain of Φ. The indices of the K_ϵ-chain are generated from a geometric distribution with parameter $1-\epsilon$. K_ϵ chains enjoy much stronger regularity than the original chain and can be used to establish many properties of the original chain.

Proposition 1 (The (weak) Markov property) *If $\mathbf{\Phi}$ is a Markov chain with initial measure μ and $h : \mathcal{X}^\infty \to \mathbb{R}$ is a bounded, measurable function, then*

$$\mathbb{E}_\mu[h(\Phi_{n+1}, \Phi_{n+2}, \dots) \mid \Phi_0, \dots, \Phi_n; \Phi_n = x] = \mathbb{E}_x[h(\Phi_1, \Phi_2, \dots)].$$

Note here that when h is the indicator function, this is just the definition of a Markov chain.

1.3.2 Occupation, Hitting and Stopping Times

The analysis of Markov chains concern its behavior (distributions) at certain random times in its evolution, which we define now.

Definition 10 (Occupation Times, Return Times and Hitting Times) *Let $\mathbf{\Phi}$ be a Markov chain and for any $A \in \mathcal{B}(\mathcal{X})$*

(i) The occupation time η_A is the number of visits by $\mathbf{\Phi}$ to A after time zero:

$$\eta_A := \sum_{n=1}^{\infty} \mathbb{I}_A\{\Phi_n\}.$$

(ii) The first return by $\mathbf{\Phi}$ to A after time zero is defined by

$$\tau_A := \min\{n \geq 1 : \Phi_n \in A\}.$$

(iii) The first hitting time on A by $\mathbf{\Phi}$ is

$$\sigma_A := \min\{n \geq 0 : \Phi_n \in A\}.$$

(iv) A function $\zeta : \mathcal{X}^\infty \to \mathbb{N}_+ \cup \{\infty\}$ is a stopping time for $\mathbf{\Phi}$ if for any initial distribution μ the event $\{\zeta = n\}$ is ν-measurable where ν is the σ-algebra induced by $\{\Phi_i\}_0^n$.

Note that η_A, τ_A and σ_A are all measurable functions from $\mathcal{X}^\infty$ to $\mathbb{N}_+ \cup \{\infty\}$ and that τ_A and σ_A are examples of stopping times.

Analysis of the stability properties of $\mathbf{\Phi}$ involves the kernel $U : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to \mathbb{R}_+ \cup \{\infty\}$, defined by

$$U(x, A) := \mathbb{E}_x(\eta_A) = \mathbb{E}_x\left(\sum_{n=1}^{\infty} \mathbb{I}_A(\Phi_n)\right) = \sum_{n=1}^{\infty} P^n(x, A), \quad x \in \mathcal{X},\, A \in \mathcal{B}(\mathcal{X})$$

and the return time probabilities

$$L(x, A) := P_x(\tau_A < \infty) = P_x(\mathbf{\Phi} \text{ ever enters } A).$$

Proposition 2 *Let $\mathbf{\Phi}$ be a Markov chain with probability transition kernel $P(x, A)$ and $n \in \mathbb{N}_+$.*

(i) For all $x \in \mathcal{X}$ and $A \in \mathcal{B}(\mathcal{X})$

$$P_x(\tau_A = 1) = P(x, A)$$

and for $n > 1$

$$\begin{aligned} P_x(\tau_A = n) &= \int_{A^c} P(x, dy) P_y(\tau_A = n-1) \\ &= \int_{A^c} P(x, dy_1) \int_{A^c} P(y_1, dy_2) \cdots \int_{A^c} P(y_{n-2}, dy_{n-1}) P(y_{n-1}, A). \end{aligned}$$

(ii) For all $x \in \mathcal{X}$ and $A \in \mathcal{B}(\mathcal{X})$

$$P_x(\sigma_A = 0) = \mathbb{I}_A(x)$$

and for $n > 0$ and $x \in A^c$

$$P_x(\sigma_A = n) = P_x(\tau_A = n).$$

Furthermore,

$$L(x, A) = P_x(\tau_A < \infty) = \sum_{n=1}^{\infty} P_x(\tau_A = n).$$

For a stopping time ζ the property that tells us that the future evolution of $\mathbf{\Phi}$ after the stopping time depends only on the value of Φ_ζ is called the strong Markov property.

Proposition 3 (The Strong Markov Property) *We say that $\mathbf{\Phi}$ has the strong Markov property if for any initial distribution μ and bounded measurable function $h : \mathcal{X}^\infty \to \mathbb{R}$ and for any stopping time ζ which is finite almost surely,*

$$\mathbb{E}_\mu[h(\Phi_{\zeta+1}, \Phi_{\zeta+2}, \dots) \mid \Phi_0, \dots \Phi_\zeta; \Phi_\zeta = x_\zeta] = \mathbb{E}_{x_\zeta}[h(\Phi_1, \Phi_2, \dots)]$$

What is the difference between the weak and strong Markov properties?