Chapter 7

Markov chain background

A stochastic process is a family of random variables $\{X_t\}$ indexed by a varaible t which we will think of as time. Time can be discrete or continuous. We will only consider the case of discrete time since that is what is relevant for MCMC. So we consider a sequence of random variables X_n indexed by a non-negative integer. The set of possible values of the random variables X_n is called the state space. It could be finite, countable or an uncountable set like R or R^d. The intuitive definition of a Markov process is that if you know $X_n = x$, then the probability distrbution of where you go next, i.e., of X_{n+1} only depends on x, not on how the process got to x. Loosely speaking, this says that the future (time $n+1$) depends on the past (times $1, 2, \cdots, n$) only through the present (time n).

We start with Markov chains with finite and then countable state spaces. For these sections I am extracting material from Durrett's *Essentials of Stochastic Processes*. Then for the section on uncountable state spaces I follow chapter 6 in the Robert and Casella book. We note that our study of Markov processes will be somewhat unbalanced since we are focusing on just the things we need for MCMC.

7.1 Finite state space

The state space is the set of possible values of the random variables X_n. In this section we study the case of finite S. A Markov chain is specified by giving a collection of transition probabilities $p(x, y)$ where $x, y \in S$. $p(x, y)$ is the probability of jumping to state y given that we are presently in state x. So if we keep x fixed and sum over y we must get 1.

Definition 1 *A transition function $p(x, y)$ is a non-negative function on $S \times S$ such that*

$$\sum_{y \in S} p(x, y) = 1 \tag{7.1}$$

To completely specify the Markov chain we also need to give the initial distribution of the chain, i.e., the distribution of X_0.

Definition 2 *Let S be a finite set and $p(x, y)$ a transition function for S. Let $\pi_0(x)$ be a probability distribuion on S. Then the Markov chain corresponding to initial distribution π_0 and transition probabilities $p(x, y)$ is the stochastic process X_n such that*

$$P(X_0 = x) = \pi_0(x), \tag{7.2}$$
$$P(X_{n+1} = x_{n+1} | X_n = x_n, X_{n-1} = x_{n-1}, \cdots, X_2 = x_2, X_1 = x_1) = p(x_n, x_{n+1}) \tag{7.3}$$

Note that the last equation implies that

$$P(X_{n+1} = x_{n+1} | X_n = x_n) = p(x_n, x_{n+1}) \tag{7.4}$$

A hard core mathematician would feel compelled to prove at this point that such a process exists. This is not that easy (it requires an extension theorem), and we will not do it. We do note that if we want to sample the process, i.e., generate a sample path for X_n, this is straightforward. Sample X_0 according to the distribution π_0. Then sample X_1 according to the distribution $p(X_0, \cdot)$. Continue, sample X_n according to the distribution $p(X_{n-1}, \cdot)$.

The following is an easy consequence of the equations above

Proposition 1 *For any states $x_0, x_1, \cdots, x_n$,*

$$P(X_0 = x_0, X_1 = x_1, \cdots, X_n = x_n) = p(x_0, x_1)p(x_1, x_2) \cdots p(x_{n-1}, x_n) \tag{7.5}$$

The transition matrix gives the transition probabilities for one time step. We consider the transition probabilities for longer times. Let $p^m(x, y) = P(X_{n+m} = y | X_n = x)$.

Proposition 2 *(Chapman-Kolmogorov equation)*

$$p^{m+k}(x, y) = \sum_z p^m(x, z)p^k(z, y) \tag{7.6}$$

If we think of $p^n(x, y)$ as the elements of a matrix, then this matrix is just p raised to the n where p is the matrix with matrix elements $p(x, y)$. If $\pi_n(x) = P(X_n = x)$ and we think of π_n as a row vector, then $\pi_n = \pi_0 p^n$.

We are primarily interested in the long time behavior of our chain. We impose some conditions on the chain for this study to rule out chains that we do not care about for MCMC. It is possible that the state space decomposes into several subsets with no transitions between the subsets. Or there could be subsets which have transitions out of the subset but not into it. **Give some examples of these** To eliminate these sorts of chains we make the following definition. The definition says that a chain is irreducible if it is possible to transition from any state to any other state in some finite number of steps.

Definition 3 *A Markov chain is irreducible if for every $x, y \in S$ there is an integer n and states $x_0, x_1, \cdots x_n$ such that $x = x_0$ and $y = x_n$ and $p(x_{i-1}, x_i) > 0$ for $i = 1, \cdots, n$. In other words, there is an integer n such that $p^n(x, y) > 0$*

Define $T_x = \min\{n \geq 1 : X_n = y\}$. This is the time of the first return (after time 0) to state x. Let P_x denote the probability measure when $X_0 = x$. A state is recurrent if $P_x(T_x < \infty) = 1$. So if we start in x we will eventually return to x. If this probability is less than 1 we say the state is transient. It can be shown that if a finite state Markov chain is irreducible, then every state x is recurrent. Finite state Markov chains can have transient states, but only if they are not irreducible.

We need to rule out one more type of chain. **Give example of periodic chain.**

Definition 4 *Let $x \in S$. The period of x is the greatest common division of the set of integers n such that $p^n(x, y) > 0$.*

Theorem 1 *In an irreducible chain all the states have the same period.*

Definition 5 *An irreducible chain is aperiodic if the common period of the states is* 1.

Note that if there is a state x such that $p(x, x) > 0$, then the period of x is 1. So if we have an irreducible chain with a state x such that $p(x, x) > 0$ then the chain is aperiodic. The condition $p(x, x) > 0$ says that if you are in state x there there is nonzero probabiity that you stay in state x for the next time step. In many applications of Markov Chains to Monte Carlo there are states with this property, and so they are aperiodic.

An irreducible, aperiodic Markov chain has nice long time behavior. It is determined by the stationary distribution.

Definition 6 *A distribution $\pi(x)$ on S is stationary if $\pi P = \pi$, i.e.,*

$$\sum_{y \in S} \pi(y) p(y, x) = \pi(x) \tag{7.7}$$

In words, a distribution is stationary if it is invariant under the time evolution. If we take the initial distribution to be the stationary distribution, than for all n the distribution of X_n is the stationary distribution. Note that the definition says that π is a left eigenvector of the transition matrix with eigenvalue 1. It may seem a little strange to be working with left eigenvectors rather than the usual right eigenvectors, but this is just a consequence of the convention that $P(X_{n+1} = y|X_n = x)$ is $p(x, y)$ rather than $p(y, x)$.

Theorem 2 *An irreducible Markov chain has a unique stationary distribution π. Furthermore, for all states x, $\pi(x) > 0$.*

Idea of proof: Eq. (7.1) implies that the constant vector is a right eigenvector of the transition matrix with eigenvalue 1. So there must exist a left eigenvector with eigenvalue 1. To see that it can be chosen to have all positive entries and is unique one can use the Perron-Frobenius theorem. Durrett has a nice proof.

Stop - Mon, 2/22

In many problems there is a short-cut for finding the stationary distribution.

Definition 7 *A state π is said to satisfy detailed balance if for all states x, y*

$$\pi(x)p(x, y) = \pi(y)p(y, x) \tag{7.8}$$

Note there are no sums in this equation.

Proposition 3 *If a distribution π satisfies detailed balance, then π is a stationary distribution.*

Proof: Sum the above equation on y. **QED**

The converse is very false. **Example**

There are two types of convergence theorems. In the first type we start the chain in some initial distribution and ask what happens to the distribution of X_n as $n \to \infty$.

Theorem 3 *Let $p(x, y)$ be the transition matrix of an irreducible, aperiodic finite state Markov chain. Then for all states x, y,*

$$\lim_{n\to\infty} p^n(x, y) = \pi(y) \tag{7.9}$$

For any initial distribution π_o, the distribution π_n of X_n converges to the stationary distribution π.

The second type of convergence theorem is not a statement about distributions. It is a statement that involves a single sample path of the process.

Theorem 4 *Consider an irreducible, finite state Markov chain. Let $f(x)$ be a function on the state space, and let π be the stationary distribution. Then for any initial distribution,*

$$P(\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(X_k) = \sum_x f(x)\pi(x)) = 1 \tag{7.10}$$

The second convergence theorem is the one that is relevant to MCMC. It says that if we want to compute the expected value of f in the probability measure π and π is the stationary distribution of some Markov chain, then we can run the chain for a long time and compute the long time average of $f(X_k)$ to get an approximation to the expected value of f.

7.2 Countable state space

Much of the finite state stuff carries over immediately. In particular the Chapman-Komogorov eq. and the fact that the n step transition matrix is the n power of the transition matrix. (Note that we now have infinite matrices.) The definition of irreducible and the period of a state is the same. And in an irreducible Markov chain, all states have the same period.

Recall that $T_x = \min\{n \geq 1 : X_n = x\}$, the time of the first return to state x. And a state is recurrent if $P_x(T_x < \infty) = 1$. There are two big changes when we go to infinite but countable state spaces. First, there can be transient states even if the chain is irreducible. Second, irreducible chains need not have stationary distibutions when they are recurrent. The definition of recurrence needs to be refined.

In a finite state Markov chain the expected value $E_x[T_x]$ is always finite for a recurrent state. But in an infinite chain, it can be infinite. If $E_x[T_x] < \infty$ we say the state is **positive recurrent.** If $E_x[T_x] = \infty$ but $P_x(T_x < \infty) = 1$, we say the state is **null recurrent.** States that are neither null or positive recurrent are said to be **transient**.

Theorem 5 *In an irreducible chain either*
(i) All states are transient
(ii) All states are null recurrent
(iii) All states are positive recurrent

MORE Example We consider a random walk on the non-negative integers. Let $0 < p < 1$. The walk jumps right with probability p, left with probability $1 - p$. If it is at the origin it jumps right with probability 1. Chain is positive recurrent if $p < 1/2$, null recurrent if $p = 1/2$ and transient if $p > 1/2$.

Theorem 6 *For an irreducible Markov chain with countable state space the following are equivalent.*

(i) All states are positive recurrent.

(ii) There is a stationary distribution π. (It is a distribution, so in particular $\sum_x \pi(x) < \infty$.)

The two big convergence theorems of the previous section hold if we add the hypothesis that there is a stationary distribution.

Theorem 7 *Let $p(x, y)$ be the transition matrix of an irreducible, aperiodic countable state Markov chain which has a stationary distribution. Then for all states x, y,*

$$\lim_{n\to\infty} p^n(x, y) = \pi(y) \tag{7.11}$$

For any initial distribution π_0, the distribution π_n of X_n converges to the stationary distribution π in the total variation norm.

In the setting of a discrete (finite or countable) state space, the total variation norm is just an l^1 norm:

$$||p^n(x, \cdot) = \pi(\cdot)||_{TV} = \sum_y |p^n(x, y) = \pi(y)| \tag{7.12}$$

Theorem 8 *Consider an irreducible, countable state Markov chain which has a stationary distribution π. Let $f(x)$ be a function on the state space such that $\sum_x |f(x)|\pi(x) < \infty$. Then for any initial distribution,*

$$P(\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(X_k) = \sum_x f(x)\pi(x)) = 1 \tag{7.13}$$

The definition of detailed balance is just the same as in the finite state space case. And we have

Proposition 4 *If a distribution π satisfies detailed balance, then π is a stationary distribution.*

7.3 General state space

For finite or countable state spaces the transition kernel is a function on $S \times S$. Since $\sum_y p(x,y) = 1$, if we fix x we can think of $p(x,\cdot)$ as a probability measure. For $A \subset S$, define

$$K(x,A) = \sum_{y \in A} p(x,y) \tag{7.14}$$

So $K(x,A)$ is the probability we jump to some state in the set A given that we are currently in state x.

Definition 8 *Let S be a set and $\mathcal{S}$ a σ-field on S. The set S is called the state space. A transition kernel K is a function from $S \times \mathcal{S}$ into $[0,1]$ such that*

(i) For all $x \in S$, $K(x,\cdot)$ is a probability measure on $(S,\mathcal{S})$.

(ii) For all $A \in \mathcal{S}$, $K(\cdot,A)$ is a measurable function on S.

If S is finite or countable, then the transition function we considered in the previous section is just given by $p(x,y) = K(x,\{y\})$

Suppose the state space is a subset of R^d and for all x, the measure $K(x,\cdot)$ is absolutely continuous with respect to Lebesgue measure. So there is a non-negative function $k(x,y)$ such that

$$K(x,A) = \int_A k(x,y)\,dy \tag{7.15}$$

for $A \in \mathcal{S}$. In this case we will refer to $k(x,y)$ as the transition function. Note that it must satisfy

$$\int_S k(x,y)dy = 1, \quad \forall x \tag{7.16}$$

Note: Robert and Casella write the density for the transition kernel as $K(x,y)$ rather then $k(x,y)$.

There are other situations in which there is a density of sorts. Something like the following will come up when we look at Gibbs samplers. To keep the notation simpler we consider two dimensions. Suppose the state space is R^2 or a subset of it. We denote out states by (x, y) and denote the Markov process by (X_n, Y_n). We consider transition kernels that describe the following. We flip a fair coin to decide whether we change the first component or the second component. If we are changing the first component then the new state (X_{n+1}, Y_{n+1}) is (X_n, Y_{n+1}) where the distribution of Y_{n+1} is absolutely continuous with respect to 1d Lebesgue measure with a density that depends on (X_n, Y_n). And if we are changing the second component, ... So we have two functions $k_1(x, y; z)$ and $k_2(x, y; z)$ such that

$$K((x_0, y_0), \cdot) = \frac{1}{2}[\delta_{x,x_0} \times k_2(x_0, y_0; y)dy + k_2(x_0, y_0; x)dx \times \delta_{y,y_0}] \tag{7.17}$$

where we let (x, y) be the variables for the measure in $K((x_0, y_0), \cdot)$.

Definition 9 *Let K be a transition matrix on the state space $(S, \mathcal{S})$. Let μ be a probability measure on $(S, \mathcal{S})$. A sequence of random variables $X_0, X_1, X_2, \cdots$ is a Markov process with transition kernel K and initial distribution μ if for all $k = 0, 1, 2, \cdots$,*

$$P(X_{k+1} \in A | X_0, X_1, \cdots, X_k) = \int_A K(X_k, dx) \tag{7.18}$$

and the distribution of X_0 is μ.

It follows immediately from the definition that

$$P(X_{k+1} \in A | X_k) = \int_A K(X_k, dx) \tag{7.19}$$

Notation: The above equation is sometimes written as (Robert and Casella do this)

$$P(X_{k+1} \in A | x_0, x_1, \cdots, x_k) = P(X_{k+1} \in A | x_k) = \int_A K(x_k, dx) \tag{7.20}$$

This should be taken to mean

$$P(X_{k+1} \in A | X_0 = x_0, X_1 = x_1, \cdots, X_k = x_k) = P(X_{k+1} \in A | X_k = x_k) = \int_A K(x_k, dx) \tag{7.21}$$

The probability measure for the process depends on the transition kernel and the initial distribution. Typically the kernel is kept fixed, but we may consider varying the initial distribution. So we let P_μ denote the probability measure for initial distribution μ. We denote

the corresponding expectation by E_μ. The fact that such a Markov process exists is quite non-trivial.

Example (Random walk) Let ξ_n be an iid sequence of RV's, and let

$$X_n = \sum_{i=1}^{n} \xi_i \tag{7.22}$$

Since $X_{n+1} = X_n + \xi_{n+1}$, $K(x, \cdot) = \mu_{\xi_{n+1}+x}$ where $\mu_{\xi_{n+1}+x}$ denotes the distribution measure of the RV $\xi_{n+1} + x$.

Example (AR(1)) Let ϵ_n be an iid sequence. For example they could be standard normal RV's. Let θ be a real constant. Then define

$$X_n = \theta X_{n-1} + \epsilon_n \tag{7.23}$$

We take $X_0 = 0$. The transition kernel is $K(x, \cdot) = \mu_{\epsilon_{n+1}+\theta x}$.

Stop - Wed, 2/24

Let $K^n(\cdot, \cdot)$ be the function on $S \times \mathcal{S}$ given by

$$K^n(x, A) = P(X_n \in A | X_0 = x) \tag{7.24}$$

These are the n step transition kernels. We now consider the Chapman Komogorov equations.

Notation remark: Let $f(x)$ be a function on X and μ a measure on X. The integral of f with respect to μ is denoted in several ways:

$$\int_X f d\mu = \int_X f(x) d\mu = \int_X f(x) d\mu(x) = \int_X f(x) \mu(dx) \tag{7.25}$$

The last one is most commonly seen in probability rather than analysis.

Proposition 5 *Let m, n be positive integers and $A \in \mathcal{S}$. Then*

$$K^{n+m}(x, A) = \int_S K^n(y, A) K^m(x, dy) \tag{7.26}$$

for $x \in S$ and $A \in \mathcal{S}$.

The finite dimensional distributions are completely determined by μ and K. Let $A_0, A_1, \cdots, A_n \in \mathcal{S}$.

$$P(X_0 \in A_0) = \int_{A_0} K(X_0, A_1)\mu(dx_0) \tag{7.27}$$

$$P(X_0 \in A_0, X_1 \in A_1, \cdots, X_n \in A_n) = \int_{A_0} d\mu(x_0) \int_{A_1} K(x_0, dx_1) \tag{7.28}$$

$$\cdots \int_{A_n} K(x_{n-1}, dx_n) \tag{7.29}$$

Theorem 9 *(weak Markov property) Let $h(x_1, x_2, \cdots)$ be a reasonable function. Then for any initial distribution and any positive integer k,*

$$E_\mu[h(X_{k+1}, X_{x+2}, \cdots)|X_0, X_1, \cdots, X_k] = E_{X_k}[h(X_1, X_2, \cdots)] \tag{7.30}$$

Note that if we take $h(x_1, \cdots) = 1_{x_1 \in A}$ then the above equation becomes the definition of a Markov process.

The definition of irreducible for discrete state space does not work for general state space.

Definition 10 *Let ϕ be a non-zero measure on the state space. A Markov chain is ϕ-irreducible if for every $A \in \mathcal{S}$ with $\phi(A) > 0$ and every $x \in S$ there is a positive integer n such that $K^n(x, A) > 0$.*

Some caution with the meaning of this def. Consider a finite chain on $\{1, \cdots, n\}$ which is irreducible. Now add one more state $n+1$ but the only new transitions are from $n+1$ to $\{1, 2, \cdots, n\}$. So there are no transitions from the original n states to state $n+1$. This is not an irreducbile chain. $n+1$ is a transient state. But if $\phi(n+1) = 0$, this new chain is ϕ-irreducible.

Example: We return to the AR(1) example. Take the ϵ_n to be standard normal. Then we can take ϕ to be Lebesgue measure on the real line. Argue the example is ϕ irreducible. Now suppose ϵ_n is uniform on $[-1, 1]$ and $\theta > 1$. Argue the chain is not ϕ irreducible.

Now suppose $\theta > 1$ and ϵ_n is uniformly distributed on $[-1, 1]$. Start the chain at $X_0 = 0$. Argue it is not ϕ irreducible where ϕ is Lebesgue measure.

Now suppose ϵ_n is absolutely continuous with respect to Lebesgue measure and the density is positive everywhere. Then it is easy to see the chain is ϕ irreducible when ϕ is Lebesgue measure.

Given a set $A \in \mathcal{S}$ we let τ_A be the time the chain first enters A, i.e.,

$$\tau_A = \inf\{n \geq 1 : X_n \in A\} \tag{7.31}$$

And we let η_A be the number of times the chain visits A, i.e.,

$$\eta_A = \sum_{n=1}^{\infty} 1_A(X_n) \tag{7.32}$$

Note that η_A can be infinite.

Definition 11 *Let X_n be a ψ irreducible Markov chain. The chain is recurrent if for all $A \in \mathcal{S}$ with $\psi(A) > 0$ we have $E_x[\eta_A] = \infty$ for all $x \in A$.*

In the discrete case if a state is recurrent, then the probability we return to the state is 1. Once we have returned the probability we will return again is still 1, and so on. So with probability one we will return infinitely many times. So $\eta_A = 1$ with probability one. The above definition is weaker than this. In particular we will have $E_x[\eta_A] = \infty$ if $P_x(\eta_A = \infty)$ is non-zero but less than 1. To rule out some pathologies we will need a stronger notion of recurrent for our convergence theorems.

Definition 12 *The chain is Harris recurrent if there is a measure ψ such that for $A \in \mathcal{S}$ with $\psi(A) > 0$ and all $x \in A$ we have $P_x[\tau_A < \infty] = 1$ for all $x \in A$.*

Definition 13 *A σ-finite measure π is invariant for a Markov chain with transition kernel K such that*

$$\pi(B) = \int_S K(x, B)\pi(dx), \quad \forall B \in \mathcal{S} \tag{7.33}$$

(Note that we do not require that it be a probability measure or even that it is a finite measure). If there is an invariant measure which is a probability measure then we say the chain is positive recurrent.

Note: Robert and Casella say just "positive" instead of "positive recurrent."

Theorem 10 *Every recurrent chain has an invariant σ-finite measure. It is unique up to a multiplicative constant.*

Example: For random walk Lebesgue measure is an invariant measure.

Example: Consider the AR(1) example when the ϵ_n have a standard normal distribution. We look for a stationary distribution with a normal distribution with mean μ and variance σ^2. If X_n is $N(\mu, \sigma^2)$ then $X_{n+1} = \theta X_n + \epsilon_n$ is $N(\theta\mu, \theta^2\sigma^2 + 1)$. So it will be stationary only if $\mu = \theta\mu$ and $\sigma^2 = \theta^2\sigma^2 + 1$. This is possible only if $|\theta| < 1$, in which case $\mu = 0$ and $\sigma^2 = 1/(1 - \theta^2)$.

Proposition 6 *If the chain is positive recurrent then it is recurrent.*

A recurrent chain that is not positive recurrent is called null recurrent.

There is an analog of detailed balance if the transition kernel is given by a density, i.e., the state space is a subset of R^d and for all x

$$K(x, A) = \int_A k(x, y)\, dy \tag{7.34}$$

for $A \in \mathcal{S}$.

Definition 14 *A chain for which the transition kernel is given by a density satisfies detailed balance if there is a non-negative function $\pi(x)$ on S such that*

$$\pi(y)k(y, x) = \pi(x)k(x, y), \quad \forall x, y \in S \tag{7.35}$$

Proposition 7 *If the chain satisfies detailed balance then π is a stationary measure.*

Proof: Integrate the detailed balance equation over y with respect to Lebesgue measure. **QED**

As we already note there will be MCMC algorithms in which the transition kernel is not given by a density but is given by a lower dimensional density. There is an analog of detailed balance in this case.

Markov chains with continuous state space can still be periodic. We give a trivial example.

Example: Let $S = [0, 1] \cup [2, 3]$. Define $K(x, \cdot)$ to be the uniform measure on $[2, 3]$ if $x \in [0, 1]$ and the uniform measure on $[0, 1]$ if $x \in [2, 3]$. Clearly if we start the chain in $[0, 1]$, then after n steps it will be somewhere in $[0, 1]$ if n is even and somewhere in $[2, 3]$ if n is odd.

The definition of period for a general state space is a bit technical and we will skip it.

As in the previous section there are two types of convergence theorems.

Theorem 11 *If the chain is Haris positive and aperiodic then for every initial distbution μ,*

$$\lim_{n\to\infty} ||K^n(x,\cdot) - \pi||_{TV} = 0 \tag{7.36}$$

where $||\quad||_{TV}$ is the total variation norm.

Explain the total variation norm

Theorem 12 *(Ergodic theorem) Suppose that the Markov chain has an invariant measure π. Then the following two statements are equivalent.*
(i) The chain is Harris recurrent.
(ii) For all $f, g \in L^1(\pi)$ with $\int_S g(x)d\pi(x) \neq 0$ we have

$$\lim_{n\to\infty} \frac{\frac{1}{n}\sum_{k=1}^n f(X_k)}{\frac{1}{n}\sum_{k=1}^n g(X_k)} = \frac{\int_S f(x)d\pi(x)}{\int_S g(x)d\pi(x)} \tag{7.37}$$

Corollary If the Markov chain is Harris recurrent and has an invariant probability measure π, then for all $f \in L^1(\pi)$ we have

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^n f(X_k) = \int_S f(x)d\pi(x) \tag{7.38}$$