Stat 8112 Lecture Notes

Markov Chains

Charles J. Geyer

April 29, 2012

1 Signed Measures and Kernels

1.1 Definitions

A *signed measure* on a measurable space $(\Omega, \mathcal{A})$ is a function $\lambda : \mathcal{A} \to \mathbb{R}$ that is countably additive, that is,

$$\lambda\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{n} \lambda(A_i),$$

whenever the sets A_i are disjoint (Rudin, 1986, Section 6.6).

A *kernel* on a measurable space $(\Omega, \mathcal{A})$ is a function $K : \Omega \times \mathcal{A} \to \mathbb{R}$ having the following properties (Nummelin, 1984, Section 1.1).

(i) For each fixed $A \in \mathcal{A}$, the function $x \mapsto K(x, A)$ is Borel measurable.

(ii) For each fixed $x \in \Omega$, the function $A \mapsto K(x, A)$ is a signed measure.

A kernel is *nonnegative* if all of its values are nonnegative. A kernel is *substochastic* if it is nonnegative and

$$K(x, \Omega) \leq 1, \qquad x \in \Omega.$$

A kernel is *stochastic* or *Markov* if it is nonnegative and

$$K(x, \Omega) = 1, \qquad x \in \Omega.$$

1.2 Operations

Signed measures and kernels have the following operations (Nummelin, 1984, Section 1.1). For any signed measure λ and kernel K, we can "left multiply" K by λ giving another signed measure, denoted $\mu = \lambda K$, defined by

$$\mu(A) = \int \lambda(dx) K(x, A), \qquad A \in \mathcal{A}.$$

For any two kernels K_1 and K_2 we can "multiply" them giving another kernel, denoted $K_3 = K_1 K_2$, defined by

$$K_3(x, A) = \int K_1(x, dy) K_2(y, A), \qquad A \in \mathcal{A}.$$

For any kernels K and measurable function $f : \Omega \to \mathbb{R}$, we can "right multiply" K by f giving another measurable function, denoted $g = Kf$, defined by

$$g(x) = \int K(x, dy) f(y), \qquad A \in \mathcal{A}, \tag{1}$$

provided the integral exists (we can only write Kf when we know the integral exists).

The kernel which acts as an identity element for kernel multiplication is defined by

$$I(x, A) = \begin{cases} 1, & x \in A \\ 0, & \text{otherwise} \end{cases}$$

Note that this combines two familiar notions. The map $x \mapsto I(x, A)$ is the indicator function of the set A, and the map $A \mapsto I(x, A)$ is the probability measure concentrated at the point x. It is easily checked that I does act as an identity element, that is $\lambda = \lambda I$ when λ is a signed measure, $KI = K = IK$ when K is a kernel, and $If = f$ when f is a bounded measurable function.

For any kernel K we write K^n for the product of K with itself n times. We also write $K^1 = K$ and $K^0 = I$, so we have $K^m K^n = K^{m+n}$ for any nonnegative integers m and n.

1.3 Finite State Space

The notation for these operations is meant to recall the notation for matrix multiplication. When Ω is a finite set, we can associate signed measures and functions with vectors and kernels with matrices and the "multiplication" operations defined above become multiplications of the associated matrices.

We take the vector space to be $\mathbb{R}^\Omega$ so vectors are functions $\Omega \to \mathbb{R}$, that is, we are taking Ω to be the index set and writing $v(x)$ rather than v_x for the components of a vector v. Then matrices are elements of $\mathbb{R}^{\Omega \times \Omega}$, so they are functions $\Omega \times \Omega \to \mathbb{R}$, that is, we are again taking Ω to be the index set and writing $m(x, y)$ rather than m_{xy} for the components of a matrix M.

We can associate a signed measure λ with a vector $\tilde{\lambda}$ defined by

$$\tilde{\lambda}(x) = \lambda(\{x\}), \qquad x \in \Omega,$$

and can associate a kernel K with a matrix $\widetilde{K}$ having elements defined by

$$\tilde{k}(x, y) = K(x, \{y\}), \qquad x, y \in \Omega.$$

A function $f : \Omega \to \mathbb{R}$ is a vector already. Think of signed measures as row vectors, then the matrix multiplication $\widetilde{\lambda}\widetilde{K}$ is associated with the kernel λK.

Think of functions as column vectors, then the matrix multiplication $\widetilde{K} f$ is associated with the function $K f$. The matrix multiplication $\widetilde{K}_1 \widetilde{K}_2$ is associated with the kernel $K_1 K_2$.

The matrix associated with the identity kernel is the identity matrix with elements $\tilde{i}(x, y) = I(x, \{y\})$.

1.4 Regular Conditional Probabilities

A Markov kernel gives a *regular conditional probability*, it describes the conditional distribution of two random variables, say of Y given X. This is often written

$$K(x, A) = \Pr(Y \in A \mid X = x), \tag{2}$$

but the right side is undefined when $\Pr(X = x) = 0$, so (2) is not really mathematics. Rather it is just an abbreviation for

$$K(\,\cdot\,, A) = P(A \mid X), \qquad \text{almost surely,} \tag{3}$$

where the right side is the measure-theoretic conditional probability defined in the handout on stationary processes.

Regular conditional probabilities do not always exist, unlike measure-theoretic conditional probabilities. That is, although there always exist random variables $P(A \mid X)$, one for each measurable set A that have the properties discussed in the handout on stationary stochastic processes, each $P(A \mid X)$ defined uniquely up to redefinition on sets of probability zero, that does not guarantee there is a kernel K that makes (3) hold for all $A \in \mathcal{A}$. Fristedt and Gray (1996, Theorem 19 of Chapter 21) give one condition, that $(\Omega, \mathcal{A})$ is a Borel space, which assures the existence and uniqueness of regular conditional probabilities.

2 Markov Chains

A stochastic process $X_1, X_2, \ldots$ taking values in an arbitrary measurable space (the X_i need not be real-valued or vector-valued), which is called the *state space* of the process, is a *Markov chain* if has the *Markov property*:

the conditional distribution of the future given the past and present depends only on the present, that is, the conditional distribution of $(X_{n+1}, X_{n+2}, \ldots)$ given $(X_1, \ldots, X_n)$ depends only on X_n. A Markov chain has *stationary transition probabilities* if the conditional distribution of X_{n+1} given X_n does not depend on n. We assume stationary transition probabilities without further mention throughout this handout.

In this handout we are interested in Markov chains on general state spaces, where "general" does not mean completely general (sorry about that), but means the measurable space $(\Omega, \mathcal{A})$ is countably generated, meaning $\mathcal{A} = \sigma(\mathcal{C})$, where $\mathcal{C}$ is a countable family of subsets of Ω. This is the assumption made by the authoritative books on general state space Markov chains (Nummelin, 1984; Meyn and Tweedie, 2009). Countably generated is a very weak assumption (it applies to the Borel sigma-algebra of any Polish space, for example). We always assume it, but will not mention it again except in Section 11 where the reason for this assumption will be explained.

We assume the conditional distribution of X_{n+1} given X_n is given by a Markov kernel P. The marginal distribution of X_1 is called the *initial distribution*. Together the initial distribution and the transition probability kernel determine the joint distribution of the stochastic process that is the Markov chain. Straightforwardly, they determine all the finite-dimensional distributions, the joint distribution of X_1, ..., X_n for any n is determined by

$$\begin{aligned} &E\{g(X_1, \ldots, X_n)\} \\ &\qquad = \int \cdots \int g(x_1, \ldots, x_n) \lambda(dx_1) P(x_1, dx_2) P(x_2, dx_3) \cdots P(x_{n-1}, dx_n), \end{aligned}$$

for all bounded measurable functions $g(X_1, \ldots, X_n)$. Fristedt and Gray (1996, Sections 22.1 and 22.3) discuss the construction of the probability measure governing the infinite sequence, showing it is determined by the finite-dimensional distributions.

For any nonnegative integer n, the kernel P^n gives the n-step transition probabilities of the Markov chain. In sloppy notation,

$$P^n(x, A) = \Pr(X_{n+1} \in A \mid X_1 = x).$$

In a different sloppy notation, we can write the joint probability measure of $(X_2, \ldots, X_{n+1})$ given X_1 as

$$P(x_1, dx_2) P(x_2, dx_3) \cdots P(x_n, dx_{n+1}),$$

which is shorthand for

$$\begin{aligned} &E\{g(X_2, \ldots, X_{n+1}) \mid X_1 = x_1\} \\ &\qquad = \int \cdots \int g(x_2, \ldots, x_{n+1}) P(x_1, dx_2) P(x_2, dx_3) \cdots P(x_n, dx_{n+1}), \end{aligned}$$

whenever $g(X_2, \ldots, X_{n+1})$ has expectation. So

$$\begin{aligned} &\Pr(X_{n+1} \in A \mid X_1 = x_1) \\ &\qquad = \int \cdots \int I_A(x_{n+1}) P(x_1, dx_2) P(x_2, dx_3) \cdots P(x_n, dx_{n+1}), \end{aligned}$$

and this does indeed equal $P^n(x_1, A)$.

Let $(\Omega, \mathcal{A})$ be the measurable space of which X_1, X_2, $\ldots$ are random elements. This is called the *state space* of the Markov chain.

3 Irreducibility

Let φ be a positive measure on the state space $(\Omega, \mathcal{A})$, meaning $\varphi(A) \geq 0$ for all $A \in \mathcal{A}$ and $\varphi(\Omega) > 0$. We say a set $A \in \mathcal{A}$ is *φ-positive* in case $\varphi(A) > 0$. A nonnegative kernel P on the the state space is *φ-irreducible* if for every $x \in \Omega$ and φ-positive $A \in \mathcal{A}$ there exists a positive integer n (which may depend on x and A) such that $P^n(x, A) > 0$. When P is φ-irreducible, we also say φ is an *irreducibility measure* for P. We say P is *irreducible* if it is φ-irreducible for some φ. We also apply these terms to Markov chains. A Markov chain is φ-irreducible (resp. irreducible) if its transition probability kernel has this property.

This definition seems quite arbitrary in that the measure φ is quite arbitrary. Note, however that φ is used only to specify a family of null sets, which are excluded from the test (we only have to find an n such that $P^n(x, A) > 0$ for A such that $\varphi(A) > 0$).

3.1 Maximal Irreducibility Measures

If a kernel is φ-irreducible for any φ, then there always exists (Nummelin, 1984, Theorem 2.4) a *maximal irreducibility measure* ψ that specifies the minimal family of null sets, meaning $\psi(A) = 0$ implies $\varphi(A) = 0$ for any irreducibility measure φ. A maximal irreducibility measure is not unique, but the family of null sets it specifies is unique.

3.2 Communicating Sets

A set $B \in \mathcal{A}$ is *φ-communicating* if for every $x \in B$ and every φ-positive $A \in \mathcal{A}$ such that $A \subset B$ there exists a positive integer n (which may depend on x and A such that $P^n(x, A) > 0$. Clearly the kernel P is φ-irreducible if and only if the whole state space is φ-communicating.

3.3 Subsampled Chains

Suppose P is a Markov kernel and q is the probability vector for a nonnegative-integer-valued random variable. Define

$$P_q(x, A) = \sum_{n=0}^{\infty} q_n P^n(x, A). \tag{4}$$

Then it is easily seen that P_q is also a Markov kernel. If X_1, X_2, $\ldots$ is a Markov chain having transition probability kernel P and N_1, N_2, $\ldots$ is an independent and identically distributed (IID) sequence of random variables having probability vector q that are also independent of X_1, X_2, $\ldots$, then X_{1+N_1}, $X_{1+N_1+N_2}$, $\ldots$ is a Markov chain having transition probability kernel P_q, which is said to be derived from the original chain by *subsampling*. If the random variables N_1, N_2, $\ldots$ are almost surely constant, that is, if the vector q has only one non-zero element, then we say the subsampling is *nonrandom*. Otherwise, we say it is *random*.

Lemma 1. *If P_q and P_r are subsampling kernels, then*

$$P_q P_r = P_{q*r},$$

where $q * r$ is the convolution of the probability vectors q and r defined by

$$(q * r)_n = \sum_{k=0}^{n} q_k r_{n-k}. \tag{5}$$

Proof.

$$\begin{aligned} (P_q P_r)(x, A) &= \sum_{k=0}^{\infty} \sum_{m=0}^{\infty} q_k r_m P^{k+m}(x, A) \\ &= \sum_{n=0}^{\infty} \sum_{k=0}^{n} q_k r_{n-k} P^n(x, A) \end{aligned}$$

(in the change of summation indices $n = k + m$ so $m = n - k$). □

Lemma 2. *Let q be a probability vector having no elements equal to zero. The following are equivalent (each implies the others).*

(a) *The set B is φ-communicating for the kernel P.*

(b) *The set B is φ-communicating for the kernel P_q.*

(c) *$P_q(x, A) > 0$ for every $x \in B$ and every φ-positive $A \subset B$.*

Proof. That (a) implies (c) implies (b) is clear. It remains only to be shown that (b) implies (a). So assume (b). Then for any $x \in B$ and $A \subset B$ such that $\varphi(A) > 0$ there exists an n such that $P_q^n(x, A) > 0$. Suppose r is another probability vector having no zero elements. It is clear from (5) that $q * r$ has no zero elements either. Let s be the n-fold convolution of q with itself, then (by mathematical induction) $P_q^n = P_s$ and s has no zero elements. Hence

$$P_s(x, A) = \sum_{n=0}^{\infty} s_n P^n(x, A) > 0, \tag{6}$$

hence some term in (6) must be nonzero, hence $P_n(x, A) > 0$ for some n, and this holding for all x and A implies (a). □

3.4 Separable Metric Spaces

Verifying irreducibility can be quite easy or exceedingly difficult. Here is a case when it is easy. An open set is said to be *connected* when it is not a union of a pair of disjoint open sets. A *basis* for a topological space is a family of open sets $\mathcal{B}$ such that every open set is a union of a subfamily of $\mathcal{B}$. A topological space is said to be *second countable* if it has a countable basis. Every separable metric space is second countable (balls having radii $1/n$ for integer n centered on points in the countable dense set form a basis).

Theorem 3. *Suppose a Markov chain with state space Ω has the following properties.*

(a) *Ω is a connected second countable topological space.*

(b) *Every nonempty open subset of Ω is φ-positive.*

(c) *Every point in Ω has a φ-communicating neighborhood.*

Then the Markov chain is φ-irreducible.

Proof. Let $\mathcal{U}$ be a countable basis for Ω, and let $\mathcal{B}$ be the family of φ-communicating elements of $\mathcal{U}$. We claim that $\mathcal{B}$ is also a countable basis for Ω. To prove this, consider an arbitrary open set W in Ω. Then for each $x \in W$, there there exists $U_x \in \mathcal{U}$ satisfying $x \in U_x \subset W$. By assumption x also has a φ-communicating neighborhood N_x whose interior N_x° contains x. Then there also exists $B_x \in \mathcal{U}$ satisfying satisfying $x \in B_x \subset U_x \cap N_x^\circ$. Since subsets of φ-communicating sets are themselves φ-communicating, $B_x \in \mathcal{B}$. This shows an arbitrary open set W is a union of elements of $\mathcal{B}$, so $\mathcal{B}$ is a basis.

Consider a sequence C_1, C_2, $\ldots$ of sets defined inductively as follows. First, C_1 is an arbitrary element of $\mathcal{B}$. Then, assuming C_1, $\ldots$, C_n have been defined, we define

$$C_{n+1} = \bigcup \{\, B \in \mathcal{B} : B \cap C_n \neq \varnothing \,\}.$$

We show each C_n is φ-communicating by mathematical induction. The base of the induction, that C_1 is φ-communicating is true by definition. To complete the induction we assume C_n is φ-communicating and must show C_{n+1} is φ-communicating. By (b) of Lemma 2 we may use a kernel P_q to show this.

So suppose $x \in C_{n+1}$ and $A \subset C_{n+1}$ is φ-positive. Because $\mathcal{B}$ is countable, there must exist $B \in \mathcal{B}$ such that $B \subset C_{n+1}$ and $\varphi(A \cap B) > 0$. Moreover we must have $B \cap C_n \neq \varnothing$ by definition of C_{n+1}. Hence $\varphi(B \cap C_n) > 0$ by assumption (b) of the theorem. Also we must have $x \in B_x$ for some $B_x \in \mathcal{B}$ such that $B_x \subset C_{n+1}$ and $B_x \cap C_n \neq \varnothing$. Hence $\varphi(B_x \cap C_n) > 0$ by assumption (b) of the theorem. Then

$$P_q^3(x, A \cap B) = \iint P_q(x, dy) P_q(y, dz) P_q(z, A \cap B)$$

is strictly positive because $P_q(z, A \cap B) > 0$ for all $z \in B$ because B is φ-communicating, and $P_q(y, B \cap C_n) > 0$ for all $y \in C_n$, because C_n is φ-communicating and $B \cap C_n$ is φ-positive, and this implies that

$$\int P_q(y, dz) P_q(z, A \cap B)$$

is strictly positive for all $y \in C_n$, and $P_q(x, B_x \cap C_n) > 0$ because B_x is φ-communicating and $B_x \cap C_n$ is φ-positive, and this implies that

$$\int P_q(x, dy) \int P_q(y, dz) P_q(z, A \cap B)$$

is strictly positive. This finishes the proof that each C_n is φ-communicating.

Let
$$C_\infty = \bigcup_{k=1}^\infty C_k.$$
Then C_∞ is φ-communicating, because any for $x \in C_\infty$ and φ-positive $A \subset C_\infty$ there is a k such that $x \in C_k$ and $\varphi(A \cap C_k) > 0$. Hence $P_q(x, A \cap C_k) > 0$ because because C_k is φ-communicating.

Now let
$$\mathcal{B}_{\text{leftovers}} = \{\, B \in \mathcal{B} : B \not\subset C_\infty \,\}.$$
Any $B \in \mathcal{B}_{\text{leftovers}}$ is actually disjoint from C_∞ because otherwise it would have to intersect some C_n and hence be contained in C_{n+1}. Thus the open set $C_{\text{leftovers}} = \bigcup \mathcal{B}_{\text{leftovers}}$ and the open set C_∞ are disjoint and their union is Ω. By the assumption that Ω is topologically connected $C_{\text{leftovers}}$ must be empty. Thus $\Omega = C_\infty$ is φ-communicating. □

The theorem seems very technical, but here is a simple toy problem that illustrates it. Let W be an arbitrary connected open subset of $\mathbb{R}^d$, and take W to be the state space of a Markov chain. Fix $\varepsilon > 0$, define $K(x, \,\cdot\,)$ to be the uniform distribution on the ball of radius ε centered at x, and define
$$P(x, A) = [1 - K(x, W)] I(x, A) + K(x, W \cap A).$$
This is a special case of the Metropolis algorithm, described in Section 9 below. The Markov chain can be described as follows. We may take X_1 to be any point of W. When the current state is X_n, we "propose" Y_n uniformly distributed on the ball of radius ε centered at X_n. Then we set
$$X_{n+1} = \begin{cases} Y_n, & Y_n \in W \\ X_n, & \text{otherwise} \end{cases} \tag{7}$$

Since W is a separable metric space it is second countable. Thus condition (a) of the theorem is satisfied. Let φ be Lebesgue measure on W. Then condition (b) of the theorem is satisfied. Let $x \in W$ and let B be the open ball of radius less than or equal to $\varepsilon/2$ centered at x and contained in W. Then $\Pr(Y_n \in B \mid X_n \in B) > 0$ and the conditional distribution of Y_n given $X_n \in B$ and $Y_n \in B$ is uniformly distributed on B. Since $Y_n \in B$ implies $Y_n \in W$ and $X_{n+1} = Y_n$, the conditional distribution of X_{n+1} given $X_n \in B$ and $Y_n \in B$ is uniformly distributed on B. This implies B is φ-communicating, and that establishes condition (c) of the theorem.

3.5 Variable at a Time Samplers

Here is another toy example that illustrates general issues. As in the example in the preceding section, we let the state space be a connected open set W in $\mathbb{R}^d$, and we show the Markov chain is φ-irreducible where φ is Lebesgue measure on W. This time, however, we use a variable at a time sampler. Fix $\varepsilon > 0$. Let $X_n(i)$ denote the i-th coordinate of the state X_n of the Markov chain (which is a d-dimensional vector). The update of the state proceeds as follows. Let I_n be uniformly distributed on the finite set $\{1, \ldots, d\}$. Let $Y_n(j) = X_n(j)$ for $j \neq I_n$, and let $Y_n(I_n)$ be uniformly distributed on the open interval $(X_n(I_n) - \varepsilon, X_n(I_n) + \varepsilon)$. Then we define X_n by (7). This is a special case of the variable-at-a-time Metropolis algorithm, which is not described in general in this handout (see Geyer, 2011).

In order to apply Theorem 3 to this example it only remains to be shown that every point of W has a φ-connected neighborhood, where φ is Lebesgue measure on W. Since W is open, every point contains a box

$$B_\delta(x) = \{\, y \in \mathbb{R}^d : |x_i - y_i| < \delta, i = 1, \ldots, d \,\}$$

such that $B_\delta(x) \subset W$ and $\delta < \varepsilon$. Fix $y \in B_\delta(x)$ and $C \subset B_\delta(x)$ such that C has positive Lebesgue measure. We claim that $P^d(y, C) > 0$. The probability that $I_k = k$, $k = 1$, $\ldots$, d is $(1/d)^d > 0$. When this occurs, we have $\Pr(Y_k \neq X_k) > (\delta/\varepsilon) > 0$, $k = 1$, $\ldots$, d. And when this occurs, we have the conditional distribution of X_d conditional on $X_d \in B_\delta(x)$ uniformly distributed on $B_\delta(x)$. Hence we have

$$P^d(y, C) \geq \left(\frac{\delta}{\varepsilon d} \right)^d \cdot \frac{\varphi(C)}{\varphi(B_\delta(x))}$$

and this is greater than zero.

3.6 Finite State Spaces

If the state space of the Markov chain is countable, then irreducibility questions can be settled by looking at paths. A *path* from x to y is a finite sequence of states

$$x = x_1, x_2, \ldots, x_n = y$$

such that

$$P(x_i, \{x_{i+1}\}) > 0, \qquad i = 1, \ldots, n - 1.$$

If there exists a state y such that there is a path from x to y for every $x \in \Omega$, then the kernel is φ-irreducible with φ concentrated at y. If there does not exist such a state y, then the kernel is not φ-irreducible for any φ.

Suppose the kernel is φ-irreducible with φ concentrated at y. Let S denote the set of states z such that there exists a path from y to z. We claim that counting measure on S is a maximal irreducibility measure ψ. Clearly, there is a path $x \to y \to z$ for any $x \in \Omega$ and $z \in S$. Thus ψ is an irreducibility measure. Conversely, if $w \notin S$, then there is no path $y \to w$. Hence no irreducibility measure can give positive measure to the point x.

4 Stationary Irreducible Markov Chains

Let P be a φ-irreducible Markov kernel. A positive measure λ is said to be *invariant* for P if it is sigma-finite and $\lambda P = \lambda$. Alternatively, we say that P *preserves* λ.

A nonnegative kernel is said to be *recurrent* if it is irreducible and has an invariant measure λ, *positive recurrent* if λ is a finite measure and *null recurrent* otherwise (this is not the definition of *recurrent* given in Meyn and Tweedie (2009), but ours is equivalent to theirs by their Theorem 10.0.1 and their Proposition 10.1.1).

Theorem 4. *If a Markov kernel is irreducible and has an invariant measure, then the invariant measure is unique up to multiplication by positive constants. Moreover, the invariant measure is a maximal irreducibility measure.*

This follows from Theorems 10.0.1 and 10.1.2 in Meyn and Tweedie (2009).

A Markov chain is *stationary* if it is a strictly stationary stochastic process. Clearly a Markov chain is stationary if it has an invariant probability measure π that is its initial distribution. Then $\pi = \pi P$ says that π is the marginal distribution of X_n for all n. Since π and P determine the finite-dimensional distributions of the Markov chain (Section 2 above), this implies the joint distribution of X_{n+1}, X_{n+2}, $\ldots$, X_{n+k} does not depend on n. Conversely, if the Markov chain is a strictly stationary stochastic process, then the marginal distribution of X_n does not depend on n, hence this marginal distribution π satisfies $\pi = \pi P$.

Stationary implies stationary transition probabilities, but not vice versa. If P is a positive recurrent Markov kernel with invariant probability measure π, then the Markov chain having initial distribution π and transition probability P is stationary. Moreover, π is the unique probability distribution having this property by Theorem 4, that is, any Markov chain having initial distribution not equal to π and transition probability P is not stationary but does have stationary transition probabilities.

5 Birkhoff Ergodic Theorem

In the context of the Birkhoff ergodic theorem we are interested in the sigma-algebra $\mathcal{S}$ of invariant events, and we are especially interested in the case where $\mathcal{S}$ is trivial, which is the same as saying the Markov chain is *ergodic* (handout on stationary stochastic processes).

Theorem 5. *Every irreducible stationary Markov chain is ergodic.*

This is Theorem 7.16 in Breiman (1968) combined with Proposition 2.3 in Nummelin (1984).

The term *functional* of a Markov chain X_1, X_2, $\dots$ refers to a time series $f(X_1)$, $f(X_2)$, $\dots$, where f is real-valued function on the state space. If X_1, X_2, $\dots$ is a stationary Markov chain, then $f(X_1)$, $f(X_2)$, $\dots$ is a strictly stationary stochastic process in the sense of the handout on stationary stochastic processes.

Note that a functional of a Markov chain is not necessarily a Markov chain: the fact that the conditional distribution of X_{n+1} given X_1, $\dots$, X_n depends only on X_n does not imply that the conditional distribution of $f(X_{n+1})$ given $f(X_1)$, $\dots$, $f(X_n)$ depends only on $f(X_n)$.

The Birkhoff ergodic theorem cannot apply to a Markov chain unless the state space is $\mathbb{R}$. Thus we are interested in sample means

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} f(X_i) \tag{8}$$

of functionals of Markov chains. We say a functional of a Markov chain is *integrable* if each of its random variables has finite expectation. For an integrable functional of a stationary irreducible Markov chain, the Birkhoff ergodic theorem says

$$\hat{\mu}_n \xrightarrow{\text{a.s.}} \mu, \tag{9}$$

where

$$\mu = E_\pi\{f(X)\} \tag{10}$$

is the expectation of $f(X)$ with respect to the (unique) invariant distribution π of the Markov chain (the statement that this is an integrable functional means the expectation exists).

But this also says something about non-stationary Markov chains having the same (stationary) transition probabilities. Let $\mathcal{X} = \Omega^\infty$ denote the sample space of the whole (infinite-dimensional) probability distribution for the whole Markov chain. Let B denote the subset of $\mathcal{X}$ consisting of points x such that

$$\hat{\mu}_n(x) = \frac{1}{n} \sum_{i=1}^{n} f(x_i) \to \mu. \tag{11}$$

That is, B^c is the null set for which the convergence promised by the Birkhoff ergodic theorem fails. Write

$$C_z = \{\, x \in B : x_1 = z \,\}$$

(the set of sequences starting with $x_1 = z$ such that (11) holds. The Fubini theorem says

$$\Pr(B) = \int \pi(dz) \Pr\{(X_1, X_2, \ldots) \in C_z \mid X_1 = z\}$$

and this can equal one only if

$$\Pr\{(X_1, X_2, X_3, \ldots) \in C_z \mid X_1 = z\} = 1, \qquad \pi \text{ almost surely.} \tag{12}$$

So this gives us a Markov chain law of large numbers conditional on the initial value. We know it holds for all initial values, except possibly a set of probability zero under the invariant distribution π, or, since π is a maximal irreducibility measure, a set of probability zero under any maximal irreducibility measure ψ. In other words, we can use any initial distribution for the Markov chain and still have (9) so long as every null set of any maximal irreducibility measure is a null set of the initial distribution.

6 Markov Chain Monte Carlo

We say h is an *unnormalized density* of a random vector X with respect to a positive measure μ if $\int h \, d\mu$ is nonzero and finite. Then the (proper, normalized, probability) density of X with respect to μ is $f = h/c$, where $c = \int h \, d\mu$.

The notion of an unnormalized density provides many Master's level probability theory homework problems of the form given h find f, but it is also very very useful in Bayesian inference and spatial statistics. Bayes rule can be phrased: likelihood times prior equals unnormalized posterior. Thus one always knows the unnormalized posterior but may not know how to normalize it. In spatial statistics and other areas of statistics involving complicated stochastic dependence amongst components of the data it is easy to specify models by unnormalized densities, because it is easy to make up functions of the data and parameters that are integrable, but it may be impossible to give closed-form expressions for those integrals and hence impossible to specify the normalized densities of the model.

The following two sections give algorithms for sampling models specified by unnormalized densities using Markov chains. One devises an irreducible Markov chain having the distribution specified by the unnormalized density as its (unique) invariant probability measure. Then one simulates the chain in a computer, considering the output X_1, X_2, $\ldots$ to be a sample from this invariant distribution in the sense that the Birkhoff ergodic theorem (9) assures that sample means of functionals of the Markov chain (8) converge to the corresponding expectations under the invariant distribution (10). This is called Markov chain Monte Carlo (MCMC).

Note that in MCMC the samples X_1, X_2, $\ldots$ are usually neither independent nor identically distributed and their distribution is not the distribution of interest (the one specified by the given unnormalized density). If the samples are independent, then they actually are IID and one is actually doing ordinary IID Monte Carlo. If the samples are identically distributed, then the Markov chain is stationary, but this is never possible to arrange unless one can produce IID samples from the distribution of interest (if one could simulate X_1 from this distribution, then why not X_2, X_3, $\ldots$?)

7 The Gibbs Sampler

The Gibbs sampler was introduced by Geman and Geman (1984) and popularized by Gelfand and Smith (1990). Why is it named after Gibbs if he didn't invent it? It was originally used to simulate Gibbs distributions in thermodynamics, which were invented by Gibbs, and it was only later realized that the algorithm applied to any distribution. Given an unnormalized density of a random vector, it may be possible to normalize the conditional distribution of each component given the other components when it is impossible to normalize the joint distribution. These conditional distributions are called the *full conditionals*.

In a *random scan* Gibbs sampler each step of the Markov chain proceeds by choosing one of the full conditionals uniformly at random and then simulating a new value of that component of the state vector using the full conditional (the remaining components do not change in this step).

In a *fixed scan* Gibbs sampler each step of the Markov chain proceeds by simulating new values of each component of the state vector using the full conditional for each (in each such simulation the remaining components do not change in that substep). The components are simulated in the same order in each step of the Markov chain. More precisely, let X_n denote the state vector and X_{ni} its components, and let f_i denote the full conditionals. Then one step of the Markov chain proceeds as follows

$$\begin{aligned}
X_{n+1,i_1} &\sim f_{i_1}(\,\cdot\, \mid X_{ni_2}, \ldots, X_{ni_d}) \\
X_{n+1,i_2} &\sim f_{i_2}(\,\cdot\, \mid X_{n+1,i_1}, X_{ni_3} \ldots, X_{ni_d}) \\
X_{n+1,i_3} &\sim f_{i_2}(\,\cdot\, \mid X_{n+1,i_1}, X_{n+1,i_2}, X_{ni_4} \ldots, X_{ni_d}) \\
&\vdots \\
X_{n+1,i_{d-1}} &\sim f_{i_{d-1}}(\,\cdot\, \mid X_{n+1,i_1}, \ldots X_{n+1,i_{d-2}}, X_{ni_d}) \\
X_{n+1,i_d} &\sim f_{i_d}(\,\cdot\, \mid X_{n+1,i_1}, \ldots X_{n+1,i_{d-1}})
\end{aligned}$$

where d is the dimension of X_n and $(i_1, \ldots, i_d)$ is a permutation of $(1, \ldots, d)$ that remains fixed for all steps of the Markov chain.

It is obvious that each substep involving the update of one coordinate preserves the distribution of interest (the one having the full conditionals being used) because if the joint distribution of all the components is the distribution of interest before the substep, then it is the same distribution afterwords (marginal times conditional equals joint). Thus a Gibbs sampler, if irreducible, simulates the distribution of interest.

8 Reversibility

A kernel K is said to be *reversible* with respect to a signed measure η if

$$\iint f(x, y) \eta(dx) K(x, dy) = \iint f(y, x) \eta(dx) K(x, dy) \tag{13}$$

for any bounded measurable function f.

The name comes from the fact that if K is a Markov kernel and η is a probability measure, then the Markov chain with transition probability kernel K and initial distribution η looks the same running forwards or backwards in time, that is, $(X_{n+1}, X_{n+2}, \ldots, X_{n+k})$ has the same distribution as $(X_{n+k}, X_{n+k-1}, \ldots, X_{n+1})$ for any positive integer k.

If a Markov kernel P is reversible with respect to a probability measure π, then π is invariant for P. To see this substitute $I_B(y)$ for $f(x, y)$ in (13), which gives

$$\begin{aligned}\int \pi(dx)P(x,B) &= \iint I_B(y)\pi(dx)P(x,dy) \\ &= \iint I_B(x)\pi(dx)P(x,dy) \\ &= \int I_B(x)\pi(dx) \\ &= \pi(B)\end{aligned}$$

which is $\pi = \pi P$.

A random scan Gibbs sampler is reversible: if the i-th component is simulated, then $X_{n+1,i}$ and X_{ni} both have the same distribution given the rest of the components (which are the same in both X_{n+1} and X_n), and this implies reversibility.

A fixed scan Gibbs sampler is not reversible (the time-reversed chain simulates components in the reverse order).

9 The Metropolis-Hastings Algorithm

Suppose h is an unnormalized density with respect to a positive measure μ on the state space and for each x in the state space $q(x,\,\cdot\,)$ is a properly normalized density with respect to μ chosen to be easy to simulate (multivariate normal, for example). The *Metropolis-Hastings algorithm* (Metropolis, Rosenbluth, Rosenbluth, Teller and Teller, 1953; Hastings, 1970) repeats the following in each step of the Markov chain.

(i) Simulate Y_n from the distribution $q(X_n,\,\cdot\,)$.

(ii) Calculate $a(X_n, Y_n)$ where

$$r(x,y) = \frac{h(y)q(y,x)}{h(x)q(x,y)} \tag{14}$$

and

$$a(x,y) = \min\bigl(1, r(x,y)\bigr). \tag{15}$$

(iii) Set $X_{n+1} = Y_n$ with probability $a(X_n, Y_n)$, and set $X_{n+1} = X_n$ with probability $1 - a(X_n, Y_n)$.

In order to avoid divide by zero in (14) it is necessary and sufficient that $h(X_1) > 0$. Proof: $q(X_n, Y_n) > 0$ with probability one because of (i), and

$h(Y_n) = 0$ implies $a(X_n, Y_n) = 0$ implies $X_{n+1} = X_n$ with probability one, hence (conversely) $X_{n+1} \neq X_n$ implies $h(X_{n+1}) > 0$.

Since the Metropolis-Hastings update is undefined when $h(X_n) = 0$, in theoretical arguments we must consider the state space to be the set of points x such that $h(x) > 0$. This is permissible, because, as was just shown, we always have $h(X_n) > 0$ even though there is no requirement that $h(Y_n) > 0$.

Terminology: Y_n is called the *proposal*, (14) is called the *Hastings ratio*, (15) is called the *acceptance probability*, substep (iii) is called *Metropolis rejection*, and the proposal is said to be *accepted* when we set $X_{n+1} = Y_n$ in step (iii) and *rejected* when we set $X_{n+1} = X_n$ in step (iii).

In the special case where $q(x, y) = q(y, x)$ for all x and y the proposal distribution q is said to be *symmetric* and this special case of the Metropolis-Hastings algorithm is called the *Metropolis algorithm.* In this special case (14) becomes

$$r(x, y) = \frac{h(y)}{h(x)} \tag{16}$$

and is called the *Metropolis ratio* or the *odds ratio.* There is little advantage to this special case. It only saves a bit of time in not having to calculate $q(X_n, Y_n)$ and $q(Y_n, X_n)$ in each step. It only gets a special name because it was proposed earlier. The Metropolis algorithm was proposed by Metropolis, et al. (1953), and the Metropolis-Hastings algorithm was proposed by Hastings (1970).

Theorem 6. *The Metropolis-Hastings update is reversible with respect to the distribution having unnormalized density h.*

Thus a Metropolis-Hastings sampler, if irreducible, simulates the distribution of interest (it does not matter what the proposal distribution is).

Proof. The kernel for the Metropolis-Hastings update is

$$P(x, A) = m(x) I(x, A) + \int_A q(x, y) a(x, y) \, \mu(dy), \tag{17}$$

where

$$m(x) = 1 - \int q(x, y) a(x, y) \, \mu(dy).$$

Let η be the measure having density h with respect to μ. Then

$$\begin{aligned}
\iint f(x,y)\eta(dx)P(x,dy) &= \iint f(x,y)h(x)P(x,dy)\mu(dx) \\
&= \iint f(x,y)m(x)I(x,dy)\mu(dx) \\
&\qquad + \iint f(x,y)h(x)q(x,y)a(x,y)\mu(dx)\mu(dy) \\
&= \int f(x,x)m(x)\mu(dx) \\
&\qquad + \iint f(x,y)h(x)q(x,y)a(x,y)\mu(dx)\mu(dy)
\end{aligned}$$

Clearly, the first term on the right side is unchanged if the arguments are interchanged in $f(x,x)$. Thus to show reversibility we only need to show that the value of

$$\iint f(x,y)h(x)q(x,y)a(x,y)\mu(dx)\mu(dy)$$

is not changed if $f(x,y)$ is changed to $f(y,x)$, and this is implied by

$$h(x)q(x,y)a(x,y) = h(y)q(y,x)a(y,x) \tag{18}$$

holding for all x and y except for those in a $\mu\times\mu$ null set. We take this null set to the the set of (x,y) such that either $h(x)q(x,y) = 0$ or $h(y)q(y,x) = 0$. For (x,y) not in this null set, we have neither the numerator nor the denominator in (14) equal to zero, and

$$r(x,y) = \frac{1}{r(y,x)}.$$

The proof of the claim (18) now splits into two cases. First, if $r(x,y) \geq 1$, so $a(x,y) = 1$, then $r(y,x) \leq 1$, so $a(y,x) = r(y,x)$, and

$$\begin{aligned}
h(y)q(y,x)a(y,x) &= h(y)q(y,x)r(y,x) \\
&= h(y)q(y,x)\frac{h(x)q(x,y)}{h(y)q(y,x)} \\
&= h(x)q(x,y) \\
&= h(x)q(x,y)a(x,y)
\end{aligned}$$

The second case is exactly the same as the first except that x and y are exchanged. □

10 Harris Recurrence

Before we even define Harris recurrence, we motivate it. We say the Markov chain law of large numbers (LLN) holds if (9) holds, where the left and right sides are given by (8) and (10). We say the Markov chain central limit theorem (CLT) holds if

$$\sqrt{n}(\hat{\mu}_n - \mu) \xrightarrow{w} \text{Normal}(0, \sigma^2) \tag{19}$$

holds, for some real number σ^2 such that $0 < \sigma^2$, where, again, the left and right sides are given by (8) and (10).

The following is Proposition 17.1.6 in Meyn and Tweedie (2009)

Theorem 7. *If the LLN (resp. CLT) holds for a stationary Harris recurrent Markov chain, then the LLN (resp. CLT) also holds for any other Markov chain having the same stationary transition probabilities.*

That is, whether the LLN or CLT holds does not depend on the initial distribution of a Harris recurrent Markov chain, so long as the chain is positive recurrent (has an invariant probability distribution) so there is a stationary Markov chain having the same transition probabilities. All initial distributions includes those concentrated at one point, so one one can also rephrase that as saying that if the LLN or CLT holds for the stationary chain, then it also holds for chains started at any point in the state space.

Now the definition: a Markov chain is *Harris recurrent* if it is irreducible with maximal irreducibility measure ψ and for every every ψ-positive set A the chain started at x hits A infinitely often with probability one. Writing this out in mathematical formulas is complicated (Meyn and Tweedie, 2009, p. 199), and we shall not do so, since one never verifies Harris recurrence directly from the definition.

For the most commonly used MCMC algorithms there are three theorems that say irreducibility implies Harris recurrence. Corollaries 1 and 2 of Tierney (1994) show this for Gibbs samplers and Metropolis-Hastings samplers that update all variables simultaneously. Theorem 1 of Chan and Geyer (1994) shows this for Metropolis-Hastings samplers that update one variable at a time (the latter requires irreducibility not only of the given Markov chain but also of all Markov chains that fix any subset of the variables).

Of course, the literature contains many other MCMC algorithms. For those one must verify Harris recurrence directly. But before we can see how to do that, we must introduce more definitions.

11 Small and Petite Sets

A subset C of the state space $(\Omega, \mathcal{A})$ is *small* if there exists a nonzero positive measure ν on the state space and a positive integer n such that

$$P^n(x, A) \geq \nu(A), \qquad A \in \mathcal{A} \text{ and } x \in C. \tag{20}$$

It is not obvious that small sets having positive irreducibility measure exist. That they do exist for any irreducible kernel P was proved by Jain and Jamison (1967) under the assumption that the state space is countably generated (this is why that assumption is always imposed).

Recall the notion of the kernel P_q derived from a kernel P by subsampling introduced in Section 3.3 above. A subset C of the state space $(\Omega, \mathcal{A})$ is *petite* if there exists a nonzero positive measure ν on the state space and a subsampling distribution q such that

$$P_q(x, A) \geq \nu(A), \qquad A \in \mathcal{A} \text{ and } x \in C. \tag{21}$$

Clearly every small set is petite (take q such that $q_n = 1$). So petite sets exist because small sets exist.

Meyn and Tweedie (2009) show that a finite union of petite sets is petite and there exists a sequence of petite sets whose union is the whole state space (their Proposition 5.5.5).

12 T-Chains

In this section we again use topology. A topological space is *locally compact* if every point has a compact neighborhood. The main example is $\mathbb{R}^d$, where for any x every closed ball centered at x is a compact neighborhood of x. Following Meyn and Tweedie (2009, Chapter 6), we assume throughout this section that the state space is a locally compact Polish space. Recall from the handout about the Wald consistency theorem that a function f on a metric space is *lower semicontinuous* (LSC) if

$$\liminf_{y \to x} f(y) \geq f(x), \qquad \text{for all } x.$$

A *continuous component* T of a kernel P having state space $(\Omega, \mathcal{A})$ is a substochastic kernel such that the function

$$x \mapsto T(x, A)$$

is LSC for any $A \in \mathcal{A}$ and there is a probability vector q such that

$$P_q(x, A) \geq T(x, A), \qquad x \in \Omega \text{ and } A \in \mathcal{A}.$$

We also say a Markov chain having P as its transition probability kernel has a continuous component T if T is a continuous component of P.

A Markov chain is a *T-chain* if it has a continuous component T such that

$$T(x, \Omega) > 0, \qquad \text{for all } x \in \Omega.$$

For a T-chain every compact set is petite and, conversely, if every compact set is petite, then the chain is a T-chain (Meyn and Tweedie, 2009, Theorem 6.0.1).

Theorem 8. *A Gibbs sampler is a T-chain if all the full conditionals are LSC functions of the variables on which they condition.*

Partial Proof. Since the notation for the Gibbs sampler is so messy, we do only the three-component case. The general idea should be clear. For both kinds of Gibbs sampler, take the continuous component T to be P itself.

For a three-component random scan Gibbs sampler, the kernel is

$$\begin{aligned} P(x, A) = \frac{1}{3} \int & I\bigl((y, x_2, x_3), A\bigr) f_1(y \mid x_2, x_3) \, dy \\ & + \frac{1}{3} \int I\bigl((x_1, y, x_3), A\bigr) f_2(y \mid x_1, x_3) \, dy \\ & + \frac{1}{3} \int I\bigl((x_1, x_2, y), A\bigr) f_3(y \mid x_1, x_2) \, dy \end{aligned}$$

and this is an LSC function of x for each fixed A by Fatou's lemma. For a three-component fixed scan Gibbs sampler that updates in the order 1, 2, 3, the kernel is

$$P(x, A) = \iiint I(y, A) f_3(y_3 \mid y_1, y_2) f_2(y_2 \mid y_1, x_3) f_1(y_1 \mid x_1, x_2) \, dy$$

and this is an LSC function of x for each fixed A by Fatou's lemma.

(For state spaces of other dimensions, the general idea is that one writes down the kernel, however messy the notation may be, and then says "and this is an LSC function of x for each fixed A by Fatou's lemma.") □

Theorem 9. *An irreducible Metropolis-Hastings sampler is a T-chain if the unnormalized density of the invariant distribution is continuous and the proposal density is separately continuous.*

Proof. As noted in Section 9 we must define the state space of the Markov chain to be the set $W = \{x : h(x) > 0\}$. The assumption that h is continuous means W is an open set.

We take the continuous component to be the part of the kernel corresponding to accepted updates, that is,

$$T(x, A) = \int_A q(x, y) a(x, y)\, dy, \tag{22}$$

where we define

$$a(x, y) = \begin{cases} 1, & h(y)q(y, x) \geq h(x)q(x, y) \\ \frac{h(y)q(y,x)}{h(x)q(x,y)}, & \text{otherwise} \end{cases}$$

(note that our definition of $a(x, y)$ avoids the problem of divide by zero when $q(x, y) = 0$, because then the first case in the definition is chosen).

Fix y and consider a sequence $x_n \to x$ with $x \in W$. It is clear that if $q(x, y) > 0$, then

$$a(x_n, y)q(x_n, y) \to a(x, y)q(x, y)$$

by the continuity assumptions of the theorem. In case $q(x, y) = 0$, we have

$$0 \leq a(x_n, y)q(x_n, y) \leq q(x_n, y) \to 0$$

by the continuity assumptions of the theorem and our definition of $a(x, y)$.

The integrand in (22) being an LSC function for each fixed value of the variable of integration, so is the integral by Fatou's lemma. It remains only to be shown that $T(x, W) > 0$ for every $x \in W$, but if this failed for any x this would mean that the chain could never move from x to anywhere and hence this chain is would not be irreducible, contrary to assumption. □

13 Periodicity

Suppose C is a small set satisfying (20) and also satisfies $\nu(C) > 0$, which is always possible to arrange (Meyn and Tweedie, 2009, Proposition 5.2.4). Define

$$E_C = \{\, n \geq 1 : (\exists \delta > 0)(\forall A \in \mathcal{A})(\forall x \in C)(P^n(x, A) \geq \delta \nu(A)) \,\}$$

Let d be the greatest common divisor of the elements of E_C. Meyn and Tweedie (2009, Theorem 5.4.4) then show that there exist disjoint measurable subsets A_0, …, A_{d-1} of the state space Ω such that

$$P(x, A_i) = 1, \qquad x \in A_j \text{ and } i = j + 1 \mod d,$$

where $j + 1 \mod d$ denotes the remainder of $j + 1$ when divided by d, and

$$\psi\big((A_0 \cup \cdots \cup A_{d-1})^c\big) = 0,$$

where ψ is a maximal irreducibility measure.

If $d \geq 2$ we say the Markov chain is *periodic* with *period d*. Otherwise, we say the Markov chain is *aperiodic*. We use the same terminology for the transition probability kernel (since whether the Markov chain is periodic or not depends only on the kernel not on the initial distribution).

For an obvious example of a periodic chain, consider a chain with state space $0, \ldots, d-1$ and deterministic movement: $X_n = x$ then $X_{n+1} = x + 1 \mod d$.

In MCMC the possibility of periodicity is mostly a nuisance. No Markov chain used in practical MCMC applications is periodic.

Theorem 10. *A positive recurrent Markov kernel of the form*

$$P(x, A) = m(x)I(x, A) + K(x, A)$$

is aperiodic if $\int m \, d\pi > 0$, where π is the invariant probability measure.

Note that (17), the kernel for a Metropolis-Hastings update has this form, where $m(x)$ is the probability that, if the current position is x, the proposal made will be rejected. In short, a Metropolis-Hastings sampler that rejects with positive probability at a set of points x having positive probability under the invariant distribution cannot be periodic.

Proof. Suppose to get a contradiction that the sampler is periodic with period d and $A_0, \ldots, A_{d-1}$ as described above. We must have $\pi(A_k) = 1/d$ for all k because $\pi(A_k) = \pi(A_{k+1 \mod d})$. Hence we have for the stationary chain

$$\Pr(X_n \in A_k \text{ and } X_{n+1} \in A_k) \geq \int_{A_k} \pi(dx) m(x)$$

and the latter is greater than zero, contradicting the periodicity assumption because A_k is π-positive. □

Theorem 11. *An irreducible Gibbs sampler is aperiodic.*

Proof. The proof begins with the same two sentences as the preceding proof. Any Gibbs update simulates X given $h_i(X)$ for some function h_i (for a traditional Gibbs sampler h_i is the projection that drops the i-th coordinate). That is, $h_i(X_{n+1}) = h_i(X_n)$ and the conditional distribution of X_{n+1} given $h_i(X_{n+1})$ is the one derived from π.

First consider a random scan Gibbs sampler. Write I_n for the random choice of which coordinate to update. Then conditional on $h_{I_n}(X_n)$ the two random elements X_n and X_{n+1} are conditionally independent. Hence

$$\Pr\bigl(X_{n+1} \in A_k \mid X_n \in A_k, h_{I_n}(X_n)\bigr) = \Pr\bigl(X_{n+1} \in A_k \mid h_{I_n}(X_n)\bigr) \tag{23}$$

In order for the sampler to be periodic, we must have

$$\begin{aligned} &\Pr(X_{n+1} \in A_k \mid X_n \in A_k) \\ &\qquad = E\bigl\{\Pr\bigl(X_{n+1} \in A_k \mid X_n \in A_k, h_{I_n}(X_n)\bigr) \mid X_n \in A_k\bigr\} \end{aligned}$$

equal to zero, and this implies (23) is zero almost surely with respect to π, but this would imply $\Pr(X_{n+1} \in A_k) = 0$, when it must be $1/d$. That is the contradiction. Since whether the chain is periodic or not does not depend on the initial distribution, this finishes the proof for random scan Gibbs samplers.

For a fixed scan Gibbs sampler, the argument is almost the same. Now there are no choices I_n and we need to consider the state between substeps. Suppose without loss of generality the scan order is 1, ..., k. Consider again the stationary chain, write $Y_0 = X_n$ and let Y_1 be the state after the first substep, Y_2, after the second, and so forth. Then conditional on $h_1(Y_0)$, $h_2(Y_1)$, ..., $h_k(Y_{k-1})$ the two random elements $X_n = Y_0$ and $X_{n+1} = Y_k$ are conditionally independent. Hence

$$\begin{aligned} &\Pr\bigl(X_{n+1} \in A_k \mid X_n \in A_k, h_1(Y_0), \dots, h_k(Y_{k-1})\bigr) \\ &\qquad = \Pr\bigl(X_{n+1} \in A_k \mid h_1(Y_0), \dots, h_k(Y_{k-1})\bigr) \end{aligned}$$

holds and contradicts the assumption of periodicity in the same way as before. Since whether the chain is periodic or not does not depend on the initial distribution, this finishes the proof for fixed scan Gibbs samplers. □

14 Total Variation Norm

The *total variation norm* of a signed measure λ on a measurable space $(\Omega, \mathcal{A})$ is defined by

$$\|\lambda\| = \sup_{A \in \mathcal{A}} \lambda(A) - \inf_{A \in \mathcal{A}} \lambda(A) \tag{24}$$

Clearly, we have

$$|\lambda(A)| \le \|\lambda\|$$

and hence

$$\sup_{A \in \mathcal{A}} |\lambda(A)| \leq \|\lambda\|.$$

Conversely,

$$\begin{aligned} \sup_{A \in \mathcal{A}} \lambda(A) &\leq \sup_{A \in \mathcal{A}} |\lambda(A)| \\ - \inf_{A \in \mathcal{A}} \lambda(A) &\leq \sup_{A \in \mathcal{A}} \bigl[- \lambda(A) \bigr] \\ &\leq \sup_{A \in \mathcal{A}} |\lambda(A)| \end{aligned}$$

so

$$\|\lambda\| \leq 2 \sup_{A \in \mathcal{A}} |\lambda(A)|.$$

In summary,

$$\sup_{A \in \mathcal{A}} |\lambda(A)| \leq \|\lambda\| \leq 2 \sup_{A \in \mathcal{A}} |\lambda(A)|.$$

For this reason one sometimes sees $\sup_{A \in \mathcal{A}} |\lambda(A)|$ referred to as the total variation norm of λ, but this does not agree with the definition used in many other areas of mathematics, which is (24).

15 The Aperiodic Ergodic Theorem

The following is Theorem 13.3.3 in Meyn and Tweedie (2009).

Theorem 12. *For a positive Harris recurrent chain with transition probability kernel P, initial distribution λ, and invariant distribution π*

$$\|\lambda P^n - \pi\| \to 0, \qquad n \to \infty.$$

This says the marginal distribution of X_n, which is λP_n, converges to π in total variation, which is a much stronger form of convergence than convergence in distribution.

Corollary 13. *For a positive Harris recurrent chain with transition probability kernel P and invariant distribution π*

$$\|P^n(x, \,\cdot\,) - \pi(\,\cdot\,)\| \to 0, \qquad n \to \infty,$$

for any x in the state space.

This is just just the special case of Theorem 12 where λ is concentrated at the point x.

16 Drift

We finally come to the high point of this handout, which is drift conditions. This section presents a drift condition for Harris recurrence. More drift conditions will come in subsequent sections.

We say a nonnegative function V is *unbounded off petite sets* if the level sets

$$\{ x \in \Omega : V(x) \le r \}$$

are petite for each real number r (Meyn and Tweedie, 2009, Section 8.4.2).

The following is Theorem 9.1.8 in Meyn and Tweedie (2009).

Theorem 14. *Suppose for an irreducible Markov chain having transition probability kernel P there exists a petite set C and a nonnegative function V that is unbounded off petite sets such that*

$$PV(x) \le V(x), \qquad x \notin C, \tag{25}$$

holds. Then the chain is Harris recurrent.

In the condition (25), the notation $PV(x)$ denotes the value of the function PV at the point x, where PV is right multiplication of the kernel P by the function V, defined by (1). It could not mean P right multiplied by $V(x)$, since it makes no sense to right-multiply a kernel by a number.

By the formulas in Section 1.2 and the interpretations of these formulas in Section 2 we can write

$$PV(x) = \int P(x, dy)V(y)$$

and interpret this as

$$PV(x) = E\{V(X_{n+1}) \mid X_n = x\}.$$

The function V is referred to as a drift function and (25) as the drift condition for recurrence.

17 Geometric Ergodicity

So far everything we have done is only enough to assure a law of large numbers (Birkhoff ergodic theorem). To get a central limit theorem, we need stronger forms of ergodicity.

The following definition is given by (Meyn and Tweedie, 2009, p. 363). A positive Harris recurrent Markov chain with transition probability kernel P and invariant distribution π is *geometrically ergodic* if there exists a real number $r > 1$ such that

$$\sum_{n=1}^{\infty} r^n \| P^n(x, \,\cdot\,) - \pi(\,\cdot\,) \| < \infty, \qquad x \in \Omega. \tag{26}$$

(Note that r does not depend on x.)

One often sees an alternative definition: a positive Harris recurrent Markov chain with transition probability kernel P and invariant distribution π is *geometrically ergodic* if there exists a real number $s < 1$ and a nonnegative function M on the state space Ω such that

$$\| P^n(x, \,\cdot\,) - \pi(\,\cdot\,) \| \le M(x) s^n, \qquad x \in \Omega. \tag{27}$$

It is obvious that (26) implies (27), but the reverse implication is almost as obvious. If we assume (27), then

$$\begin{aligned} \sum_{n=1}^{\infty} r^n \| P^n(x, \,\cdot\,) - \pi(\,\cdot\,) \| &\le \sum_{n=1}^{\infty} r^n M(x) s^n \\ &\le \frac{M(x)}{1 - rs} \end{aligned}$$

so long as $rs < 1$, and this proves (26) for any r such that $1 < r < 1/s$.

18 Geometric Drift

Recall the definition of "unbounded off petite sets" from Section 16. The following is part of Theorem 15.0.1 in Meyn and Tweedie (2009).

Theorem 15. *Suppose for an irreducible, aperiodic Markov chain having transition probability kernel P and state space Ω there exists a petite set C, a real-valued function V satisfying $V \ge 1$, and constants $b < \infty$ and $\lambda < 1$ such that*

$$PV(x) \le \lambda V(x) + b I(x, C), \qquad x \in \Omega, \tag{28}$$

holds. Then the chain is geometrically ergodic.

The function V is referred to as a drift function and (28) as the drift condition for geometric ergodicity.

Theorem 15 has a near converse, which is another part of Theorem 15.0.1 in Meyn and Tweedie (2009).

Theorem 16. *For an geometrically ergodic Markov chain having transition probability kernel P, invariant distribution π, and state space Ω, there exists an extended-real-valued function V satisfying $V \geq 1$ and $\pi\bigl(V(x) < \infty\bigr) = 1$, constants $b < \infty$ and $\lambda < 1$, and a petite set C such that* (28) *holds. Moreover, there exist constants $r > 1$ and $R < \infty$ such that*

$$\sum_{n=1}^{\infty} r^n \| P^n(x, \,\cdot\,) - \pi(\,\cdot\,) \| \leq R V(x), \qquad x \in \Omega.$$

This shows that the function M in (27) can be taken to be a positive multiple of a drift function V. Taking expectations with respect to π of both sides of (28) and using $\pi = \pi P$, we get

$$(1 - \lambda) E_\pi\{V(X)\} \leq b \pi(C),$$

which shows that a function satisfying the geometric drift condition is always π-integrable. Thus we can always take the function M in (27) to be π-integrable.

The fact that any solution V to the geometric drift condition is π-integrable gives us a way to find at least some unbounded π-integrable functions: any random variable $g(X)$ satisfying $|g| \leq V$ has expectation with respect to π.

There is an alternate form of the geometric drift condition that is often easier to verify (Meyn and Tweedie, 2009, Lemma 15.2.8).

Theorem 17. *The geometric drift condition* (28) *holds if and only if V is unbounded off petite sets and there exists a constant $L < \infty$ such that*

$$PV \leq \lambda V + L. \tag{29}$$

18.1 Example: AR(1)

An AR(1) time series defined by

$$X_{n+1} = \rho X_n + \sigma Y_n,$$

where Y_1, Y_2, $\ldots$ are IID standard normal, is a Markov chain. For homework we proved that Normal$(0, \tau^2)$ is an invariant distribution, where

$$\tau^2 = \frac{\sigma^2}{1 - \rho^2}$$

provided $\rho^2 < 1$. Clearly this Markov chain is irreducible, Lebesgue measure being an irreducibility measure, because the conditional distribution of X_{n+1} given X_n gives positive probability to every set having positive Lebesgue measure. Thus we now know this invariant distribution is unique (which we had to prove in the homework using characteristic functions).

We also showed in homework (the same proof using characteristic functions) that there does not exist an invariant probability measure unless $\rho^2 < 1$. Thus when $\rho \geq 1$, the AR(1) process is still a Markov chain but not a positive recurrent Markov chain. It can be shown that the chain is null recurrent when $\rho^2 = 1$ and transient otherwise, but we will not bother with this.

Here we show that an AR(1) process with $\rho^2 < 1$ is geometrically ergodic. First it is a T-chain because the conditional probability density function for X_{n+1} given X_n is a continuous function of X_n. Thus every compact set is petite and the function V defined by $V(x) = 1 + x^2$ is unbounded off petite sets. Now

$$PV(x) = E(1 + X_{n+1}^2 \mid X_n = x) = 1 + \rho^2 x^2 + \sigma^2$$

and hence we have the alternate geometric drift condition (29) with $\lambda = \rho^2$ and $L = 1 - \rho^2 + \sigma^2$.

18.2 A Gibbs Sampler

Suppose X_1, $\ldots$, X_n are IID Normal(μ, λ^{-1}) and we suppose that the prior distribution on (μ, λ) is the improper prior with density with respect to Lebesgue measure

$$g(\mu, \lambda) = \lambda^{-1/2}.$$

We wish to use a Gibbs sampler to simulate this (actually the joint posterior distribution can be derived in closed form, but for this example we ignore that).

The unnormalized posterior is

$$\lambda^{n/2} \exp\left(-\frac{\lambda}{2} \sum_{i=1}^{n} (x_i - \mu)^2 \right) \cdot \lambda^{-1/2}$$
$$= \lambda^{(n-1)/2} \exp\left(-\frac{n\lambda}{2} \left[v_n + (\bar{x}_n - \mu)^2 \right] \right)$$

where

$$\begin{aligned}\bar{x}_n &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ v_n &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2\end{aligned}$$

Hence the posterior conditional distribution of λ given μ is Gamma(a, b), where

$$\begin{aligned}a &= (n+1)/2 \\ b &= n\bigl[v_n + (\bar{x}_n - \mu)^2\bigr]/2\end{aligned}$$

and the posterior conditional distribution of μ given λ is Normal(c, d) where

$$\begin{aligned}c &= \bar{x}_n \\ d &= n^{-1}\lambda^{-1}\end{aligned}$$

We use a fixed scan Gibbs sampler updating first λ then μ in each iteration, that is, we simulate the Markov chain (μ_n, λ_n), $n = 1,\ 2,\ \ldots$ as follows

$$\begin{aligned}\lambda_{n+1} &\sim f_{\lambda|\mu}(\,\cdot \mid \mu_n) \\ \mu_{n+1} &\sim f_{\mu|\lambda}(\,\cdot \mid \lambda_{n+1})\end{aligned}$$

where $\sim$ means "is simulated from the distribution."

Again, we know the conditional distributions are continuous functions of the conditioning variables so the chain is a T-chain and every compact set is petite.

We try a drift function

$$V(\mu, \lambda) = 1 + (\mu - \bar{x}_n)^2 + \varepsilon\lambda^{-1} + \lambda$$

where $\varepsilon > 0$ is a constant to be named later.

Clearly, this is unbounded off compact sets of the state space which is $\mathbb{R} \times (0, \infty)$. The term $\varepsilon\lambda^{-1}$ makes $V(\mu, \lambda)$ go to infinity as λ goes to zero.

First

$$\begin{aligned}E\{V(\mu_{n+1}, \lambda_{n+1}) \mid \lambda_{n+1}, \mu_n, \lambda_n\} &= E\{V(\mu_{n+1}, \lambda_{n+1}) \mid \lambda_{n+1}\} \\ &= 1 + n^{-1}\lambda_{n+1}^{-1} + \varepsilon\lambda_{n+1}^{-1} + \lambda_{n+1} \\ &= 1 + (\varepsilon + n^{-1})\lambda_{n+1}^{-1} + \lambda_{n+1}\end{aligned}$$

so, using the facts that if X is Gamma(a, b) then

$$
\begin{aligned}
E(X^{-1}) &= \frac{b}{a-1} \\
E(X) &= \frac{a}{b}
\end{aligned}
$$

(the first requires $a - 1 > 0$, otherwise the expectation does not exist), we obtain

$$
\begin{aligned}
E\{V(\mu_{n+1}, \lambda_{n+1}) \mid \mu_n, \lambda_n\} &= E\{E[V(\mu_{n+1}, \lambda_{n+1}) \mid \lambda_{n+1}, \mu_n, \lambda_n] \mid \mu_n, \lambda_n\} \\
&= 1 + (\varepsilon + n^{-1}) E(\lambda_{n+1}^{-1} \mid \mu_n) + E(\lambda_{n+1} \mid \mu_n) \\
&= 1 + \frac{(\varepsilon + n^{-1}) n \bigl[v_n + (\bar{x}_n - \mu_n)^2\bigr]/2}{(n+1)/2 - 1} \\
&\qquad + \frac{(n+1)/2}{n \bigl[v_n + (\bar{x}_n - \mu_n)^2\bigr]/2} \\
&\le 1 + \frac{(n\varepsilon + 1) \bigl[v_n + (\bar{x}_n - \mu_n)^2\bigr]}{n-1} + \frac{n+1}{n v_n} \\
&= 1 + \frac{(n\varepsilon + 1) v_n}{n-1} + \frac{n+1}{n v_n} + \frac{(n\varepsilon + 1)(\bar{x}_n - \mu_n)^2}{n-1} \\
&\le \rho V(\mu_n, \lambda_n) + L,
\end{aligned}
$$

where

$$
\begin{aligned}
\rho &= \frac{n\varepsilon + 1}{n-1} \\
L &= 1 + \frac{(n\varepsilon + 1) v_n}{n-1} + \frac{n+1}{n v_n}
\end{aligned}
$$

Thus we satisfy the geometric drift condition if we can make $\rho < 1$, which we can if $n > 2$ and $\varepsilon < 1/n$.

19 Uniform Ergodicity

Our final and strongest form of ergodicity is this. A positive Harris recurrent Markov chain with transition probability kernel P, invariant distribution π, and state space Ω is *uniformly ergodic* if one can use a constant function M in the alternative definition of geometrically ergodic, that is, if there exist a real numbers $s < 1$ and $M < \infty$ such that

$$
\| P^n(x, \,\cdot\,) - \pi(\,\cdot\,) \| \le M s^n, \qquad x \in \Omega. \tag{30}
$$

It is obvious that uniform ergodicity (30) implies geometric ergodicity (27), but the reverse implication does not necessarily hold.

Since uniform ergodicity implies geometric ergodicity, by Theorem 16 the geometric drift condition (28) holds and the drift function V is unbounded off petite sets and the function M in the alternate definition of geometric ergodicity (27) can be taken proportional to V. But this says V must be both bounded and unbounded off petite sets, which can happen only if the whole state space is petite. Actually something stronger is true (Meyn and Tweedie, 2009, Theorem 16.0.2).

Theorem 18. *A Markov chain is uniformly ergodic if and only if the whole state space is small.*

This shows that most Markov chains are not uniformly ergodic, but here are a few cases where one does have uniform ergodicity.

- The Markov chain is actually an IID sequence, in which case it is obvious from the definition that the whole state space is a small set because $P(x, A) = \pi(A)$ for all x by independence.

- The Markov chain is a T-chain and the state space is compact.

- The state space is finite. In this case the Markov chain is automatically a T-chain because every function on a discrete topological space is continuous and the state space is automatically compact because every finite set of any topological space is compact.

In all of these cases the Markov chain is uniformly ergodic if it is positive Harris recurrent.

20 Types of Ergodicity and Mixing Coefficients

Mixing coefficients, which were introduced in the handout on stationary stochastic processes are much simplified when the stochastic process is a stationary positive Harris recurrent Markov chain (Bradley, 1986, Section 4). Let $(\Omega, \mathcal{A})$ be the state space, P the transition probability kernel, π the invariant probability distribution, and $L_2(\pi)$ the set of all functions f such that $\int f^2 \, d\pi < \infty$. In this case the mixing coefficients become

$$\alpha_n = \sup_{A, B \in \mathcal{A}} \left| \int_A \pi(dx) P^n(x, B) - \pi(A)\pi(B) \right| \tag{31}$$

$$\beta_n = \sup_{\substack{m \in \mathbb{N} \\ A_1, \ldots, A_I \in \mathcal{A} \\ A_1, \ldots, A_I \text{ partition } \Omega \\ B_1 \ldots B_J \in \mathcal{A} \\ B_1, \ldots, B_J \text{ partition } \Omega}} \frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{J} \left| \int_{A_i} \pi(dx) P^n(x, B_j) - \pi(A_i)\pi(B_j) \right| \tag{32}$$

$$\rho_n = \sup_{f, g \in L_2(\pi)} \left| \operatorname{cor}\bigl(f(X_k), g(X_{k+n})\bigr) \right| \tag{33}$$

$$\phi_n = \sup_{\substack{A, B \in \mathcal{A} \\ \pi(B) > 0}} \left| \frac{\int_A \pi(dx) P^n(x, B)}{\pi(B)} - \pi(A) \right| \tag{34}$$

Moreover, we have the following (Bradley, 1986, Theorem 4.2)

Theorem 19. *If a stationary Markov chain is rho-mixing, then it is rho-mixing exponentially fast, meaning $\rho_n \le M s^n$ for some $s < 1$ and $M < \infty$. If a stationary Markov chain is phi-mixing, then it is phi-mixing exponentially fast.*

Corollary 20. *For every square-integrable functional of a stationary, positive Harris recurrent Markov chain that is rho-mixing or phi-mixing satisfies the* (19) *with*

$$\sigma^2 = \lim_{n \to \infty} \operatorname{var}\{f(X_1) + \cdots + f(X_n)\}. \tag{35}$$

This is Theorem 19 combined with part (ii) of Theorem 6 in the handout on stationary stochastic processes combined with phi-mixing implies rho-mixing.

Theorem 21. *Every positive Harris recurrent, aperiodic, stationary Markov chain is beta-mixing, hence alpha-mixing. If such a chain is geometrically ergodic, then it is beta-mixing exponentially fast, hence alpha-mixing exponentially fast.*

Proof. The first sentence is Theorem 4.3 in Bradley (1986). The alpha-mixing part of the second sentence is in the proof of Theorem 2 in Chan and Geyer (1994), and that method of proof easily extends to beta-mixing. □

Corollary 22. *A geometrically ergodic, stationary Markov chain has a central limit theorem for all functionals having $2 + \varepsilon$ moments for some $\varepsilon > 0$,* *that is,* (19) *holds with*

$$\sigma^2 = \operatorname{var}\{f(X_n)\} + 2 \sum_{k=1}^{\infty} \operatorname{cov}\{f(X_n), f(X_{n+k})\} \tag{36}$$

This is Theorem 21 combined with part (i) of Theorem 5 in the handout on stationary stochastic processes, the formula for σ^2 coming from the latter.

For reversible Markov chains (and recall that the Metropolis-Hastings algorithm and random-scan Gibbs samplers are reversible) we can do better (Roberts and Rosenthal, 1997, Theorem 2.1).

Theorem 23. *Every reversible, geometrically ergodic, stationary Markov chain is rho-mixing.*

Corollary 24. *A reversible, geometrically ergodic, stationary Markov chain has a central limit theorem for all square-integrable functionals, that is,* (19) *holds with σ^2 given by* (36).

Theorem 25. *Every uniformly ergodic, stationary Markov chain is phi-mixing.*

Proof. This is shown on pp. 365–366 in Ibragimov and Linnik (1971). □

Corollary 26. *A uniformly ergodic, stationary Markov chain has a central limit theorem for all square-integrable functionals, that is,* (19) *holds with σ^2 given by* (35).

Lastly, recall that Theorem 7 says that if the chain is Harris recurrent all of the CLT results hold for any initial distribution, not just for stationary chains.

References

Bradley, R. C. (1986). Basic properties of strong mixing conditions. In Eberlein, E. and Taqqu, M. S., eds. *Dependence in Probability and Statistics: A Survey of Recent Results* (Oberwolfach, 1985). Boston: Birkhäuser.

Breiman, L. (1968). *Probability.* Redding, MA: Addison-Wesley. Republished 1992, Philadelphia: Society for Industrial and Applied Mathematics.

Chan, K. S. and Geyer, C. J. (1994). Discussion of Tierney (1994). *Annals of Statistics*, **22**, 1747–1758.

Fristedt, B. E. and Gray, L. F. (1996). *A Modern Approach to Probability Theory.* Boston: Birkhäuser.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398–409.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Geyer, C. J. (2011). Introduction to Markov chain Monte Carlo. In *Handbook of Markov Chain Monte Carlo*, edited by Brooks, S., Gelman, A., Jones, G., and Meng, X.-L.. Boca Raton, FL: Chapman & Hall/CRC.

Ibragimov, I. A. and Linnik, Yu. V. (1971). *Independent and Stationary Sequences of Random Variables* (edited by J. F. C. Kingman). Groningen: Wolters-Noordhoff.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Jain, N. and Jamison, B. (1967). Contributions to Doeblin's theory of Markov processes. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **8**, 19-40.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–1092.

Meyn, S. P. and Tweedie, R. L. (2009). *Markov Chains and Stochastic Stability*, second edition. Cambridge: Cambridge University Press.

Nummelin, E. (1984). *General Irreducible Markov Chains and Non-Negative Operators.* Cambridge: Cambridge University Press.

Roberts, G. O. and Rosenthal, J. S. (1997). *Electronic Communications in Probability*, **2**, 13–25.

Rudin, W. (1986). *Real and Complex Analysis*, third edition. New York: McGraw-Hill.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics*, **22**, 1701–1762.