Stat 8501 Lecture Notes
Markov Chains
Charles J. Geyer
February 26, 2020

Contents

1 Kernels

1.1 Definitions

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 measurable.

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

We will only be interested in kernels such that $K(x, \cdot)$ is a signed measure for each x.

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

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

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

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

If K is a stochastic kernel then $K(x, \cdot)$ a probability measure for each x.

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 $\tilde{\lambda}\widetilde{K}$ is associated with the kernel λK.

Think of functions as column vectors, then the matrix multiplication $\widetilde{K}f$ is associated with the function Kf. The matrix multiplication $\widetilde{K}_1\widetilde{K}_2$ is associated with the kernel K_1K_2.

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

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 the right hand side is not really mathematics. Kernels are real mathematics.

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 and $\sigma(\mathcal{C})$ is the smallest sigma-algebra containing $\mathcal{C}$, which is the intersection of all sigma-algebras containing $\mathcal{C}$. 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 $\mathbb{R}^d$, for example, the Borel sigma-algebra being $\sigma(\mathcal{O})$, where $\mathcal{O}$ is the family of open subsets of $\mathbb{R}^d$). We always assume it, but will not mention it again except in Section 6.1 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 $X_1, \ldots, X_n$ for any n is determined by

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

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

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

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

3 Irreducibility

Let φ be a strictly 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

$$P_q(x, A) = \sum_{n=0}^{\infty} q_n P^n(x, A). \tag{3}$$

Then it is easily seen that P_q is also a Markov kernel. If X_1, X_2, ... is a Markov chain having transition probability kernel P and N_1, N_2, ... is an independent and identically distributed (IID) sequence of random variables having probability vector q that are also independent of X_1, X_2, ..., then X_{1+N_1}, $X_{1+N_1+N_2}$, ... 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, ... 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

$$(q * r)_n = \sum_{k=0}^{n} q_k r_{n-k}. \tag{4}$$

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 (4) 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

$$P_s(x, A) = \sum_{n=0}^{\infty} s_n P^n(x, A) > 0, \tag{5}$$

hence some term in (5) must be nonzero, hence $P^n(x, A) > 0$ for some n, and this holding for all x and all φ-positive 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, ... of sets defined inductively as follows. First, C_1 is an arbitrary element of $\mathcal{B}$. Then, assuming C_1, ..., 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 5.5 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

$$X_{n+1} = \begin{cases} Y_n, & Y_n \in W \\ X_n, & \text{otherwise} \end{cases} \tag{6}$$

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 (6). 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 $\Pr(X_{k+1} \neq X_k) > (\delta/\varepsilon) > 0$, $k = 1, \ldots, d$. 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 Countable 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 w.

4 Stochastic Stability

4.1 Transience and Recurrence

Let N_A denote the number of visits to A made by a Markov chain (this is a random variable, different for each realization of the chain). A set A is called *recurrent* if

$$E_x(N_A) = \infty, \qquad x \in A.$$

A set A is called *uniformly transient* if there exists a real number M such that

$$E_x(N_A) \le M, \qquad x \in A.$$

For a countable state space Markov chain and A a singleton set, these definitions reduce to the usual ones (Hoel, et al., 1986, Theorem 1 of Chapter 1). But for uncountable state spaces, especially for continuous distributions, looking at single points makes no sense. If the distribution of the Markov chain is continuous, then every point has probability zero and $E_x(N_{\{y\}}) = 0$, for every point y.

For a ψ-irreducible Markov chain the behavior is simple. Either every ψ-positive set is recurrent, in which case we say the chain is *recurrent*, or there exists a countable cover of the state space by uniformly transient sets, in which case we say the chain is *transient* (Meyn and Tweedie, 2009, Theorem 8.0.1). This is the *transience-recurrence dichotomy:* every irreducible Markov chain is either transient or recurrent.

For a transient chain we can say more about which sets are uniformly transient, but this gets us a bit ahead of ourselves. Petite sets are defined in Section 6.1 below, and an easily used criterion for petiteness is given in Section 6.2 below (every compact set is petite under certain regularity conditions that hold for many Markov chains, and every subset of a petite set is petite, so for such chains every bounded set is petite). Theorem 8.0.1 in Meyn and Tweedie (2009) adds that for a transient chain every petite set is uniformly transient.

The definitions of transience and recurrence work in opposite directions. For recurrence, we might be concerned that a set is too little to be hit infinitely often, but the theorem cited above says no ψ-positive set, no matter how little, fails to be recurrent. For transience, we might be concerned that a set is too big to be hit only finitely many times, but the theorem cited above says every petite set, no matter how big, is uniformly transient.

We will learn more about transience and recurrence in the next section and also in Sections 6.6 and 6.7 below.

4.2 Subinvariant and Invariant Measures

A measure is said to be sigma-finite (also written σ-finite) if there is a countable partition of the state space such that the measure of each element of the partition is finite. A measure is said to be strictly positive if it is positive and not the zero measure.

For every irreducible Markov kernel P there exists a strictly positive, sigma-finite measure μ such that $\mu \geq \mu P$, meaning

$$\mu(A) \geq \int \mu(dx) P(x, A), \qquad A \in \mathcal{A}.$$

where $(\Omega, \mathcal{A})$ is the state space (Meyn and Tweedie, 2009, Theorem 8.0.1, Theorem 10.0.1, and Proposition 10.1.3). Such a measure μ is called *subinvariant*.

If a subinvariant measure actually satisfies $\mu = \mu P$, then it is called *invariant*. If a subinvariant measure is not invariant, then it is called *strictly subinvariant*.

If P is irreducible and recurrent, then every subinvariant measure is actually invariant and is unique up to multiplication by a positive scalar (Meyn and Tweedie, 2009, Theorem 10.4.4). If the invariant measure is finite, in which case it can be renormalized to be a probability measure, then we say the chain is *positive recurrent*. Otherwise, we say the chain is *null recurrent*. If P is irreducible, then it is positive recurrent if and only if it has a finite invariant probability measure (Meyn and Tweedie, 2009, Theorems 10.1.1 and 10.4.4).

If P is irreducible and transient, then P has a strictly subinvariant measure, which need not be not unique (Meyn and Tweedie, 2009, Proposition 10.1.3). It may or may not have invariant measures, and the invariant measures, if they exist, may or may not be unique. An invariant measure, if it exists cannot be finite (otherwise the chain would be positive recurrent).

Thus we can use invariant and subinvariant measures to classify irreducible Markov chains. The chain is positive recurrent if and only if an invariant probability measure exists. The chain is transient if and only if a strictly subinvariant measure exists. The left over case is null recurrent.

Another useful fact is that any subinvariant measure is a maximal irreducibility measure (Meyn and Tweedie, 2009, Proposition 10.1.2). This takes some of the mystery out of maximal irreducibility measures.

4.2.1 Example: AR(1) Time Series

An AR(1) time series is defined by

$$X_n = \rho X_{n-1} + \sigma Y_n, \tag{7}$$

where Y_1, Y_2, ... are IID standard normal and independent of X_0. It is clear that the conditional distribution of X_n given the past history only depends on X_{n-1}, so this is a Markov chain.

The AR in the name stands for auto-regressive, the idea being that (7) looks something like the specification for simple linear regression except that the series is being regressed on itself (the X_i play the role of response on the left hand side and the role of predictor on the right hand side). The 1 in the name indicates that there is just one "predictor." An AR(k) time series for $k > 1$ has additional terms $\rho_2 X_{n-2}$, $\rho_3 X_{n-3}$, and so forth. But these are not Markov chains, so we ignore them (they are discussed in the time series class).

By independence of X_{n-1} and Y_n we have

$$\operatorname{var}(X_n) = \rho^2 \operatorname{var}(X_{n-1}) + \sigma^2.$$

Suppose an AR(1) time series is a stationary Markov chain. Then we have $\operatorname{var}(X_n) = \operatorname{var}(X_{n-1})$. Hence

$$\operatorname{var}(X_n) = \frac{\sigma^2}{1 - \rho^2} \tag{8}$$

provided $\rho^2 < 1$ (otherwise, the right hand side clearly cannot define a variance). In case (8) does define a variance, call it τ^2.

We guess that a normal distribution is invariant and we check that. Clearly, if $E(X_{n-1}) = 0$, then $E(X_n) = 0$, too. This, together with our derivation of (8) and the fact that the sum of independent normal random variables is normal, tells us that Normal$(0, \tau^2)$ is an invariant distribution when $\rho^2 < 1$.

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 Normal$(0, \tau^2)$ is the unique invariant probability distribution and the Markov chain is positive recurrent (when $\rho^2 < 1$).

It is also fairly clear that λP is proportional to λ when λ is Lebesgue measure by symmetry of the normal distribution. Let's calculate that. For

any Lebesgue measurable set A,

$$\begin{aligned}
(\lambda P)(A) &= \int_{-\infty}^{\infty} P(x, A)\, dx \\
&= \int_{-\infty}^{\infty} dx \int_{\rho x + \sigma y \in A} \phi(y)\, dy \\
&= \int_{-\infty}^{\infty} \phi(y)\, dy \int_{\rho x + \sigma y \in A} dx \\
&= \lambda(\rho^{-1} A) \int_{-\infty}^{\infty} \phi(y)\, dy \\
&= \lambda(\rho^{-1} A) \\
&= |\rho|^{-1} \lambda(A)
\end{aligned}$$

where ϕ is the PDF of the standard normal distribution, where in the last three lines we are assuming $\rho \neq 0$, and where $\rho^{-1} A$ means multiplying every point of A by ρ^{-1}. The third equality is interchanging the order of integration, which is also valid in measure theory (this is called the Fubini theorem or the Tonelli theorem). The last equality is translation invariance of Lebesgue measure.

In case $\rho^2 < 1$ so $|\rho|^{-1} > 1$ we do not find that Lebesgue measure is subinvariant. In case $\rho^2 = 1$ so $|\rho|^{-1} = 1$ we find that Lebesgue measure is invariant. In case $\rho^2 > 1$ so $|\rho|^{-1} < 1$ we find that Lebesgue measure is strictly subinvariant. Hence we have found that the Markov chain is transient in case $\rho^2 > 1$. In case $\rho^2 = 1$ we have found that there is an invariant measure that is not a finite measure, so we have ruled out positive recurrence, but we still do not know whether the chain is transient or null recurrent.

We will settle this open issue and find out more about AR(1) Markov chains in Sections 6.7.1 and 6.8.1 below.

4.3 Stationary Markov Chains

A stochastic process is *strictly stationary* if the distribution of a block of consecutive random variables only depends on the length of the block, that is, the distribution of $X_{n+1}, \ldots, X_{n+k}$ depends only on k (does not depend on n). A Markov chain is *stationary* if it is a strictly stationary stochastic process.

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.

4.4 Harris Recurrence

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.

We will learn more about Harris recurrence in Sections 5.8 and 6.7 below.

4.5 The Law of Large Numbers

The strong law of large numbers (LLN) for IID sequences of random variables says the following. Let X_1, X_2, $\ldots$ be a sequence of IID random variables having expectation μ, and define

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i, \tag{9}$$

then

$$\overline{X}_n \xrightarrow{\text{a.s.}} \mu, \tag{10}$$

this being known as the law of large numbers.

We want to discuss the LLN for Markov chains, but if X_1, X_2, $\ldots$ is a Markov chain, the X_1 need not be real-valued, so expectation need not even be defined. Hence we introduce the notion of functionals of Markov chains.

Suppose X_1, X_2, $\ldots$ is a Markov chain and f is a real-valued function on the state space of the Markov chain, then we say the stochastic process $f(X_1)$, $f(X_2)$, $\ldots$ is a *functional* of this Markov chain.

Suppose X_1, X_2, $\ldots$ is a positive Harris recurrent Markov chain, π is its unique invariant distribution, and f is a real-valued function on the state space such that

$$\mu = E_\pi\{f(X)\} = \int f(x)\, \pi(dx) \tag{11}$$

exists. Define

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} f(X_i) \tag{12}$$

then

$$\hat{\mu}_n \xrightarrow{\text{a.s.}} \mu \tag{13}$$

(Meyn and Tweedie, 2009, Proposition 17.1.7), this being known as the Markov chain law of large numbers.

The similarity of (10) and (13) is made clearer if we define $Y_i = f(X_i)$ so the left hand side of (13) is the sample mean $\overline{Y}_n$. There is a difference that in (13) μ is not the expectation of the Y_i (indeed, if the Markov chain is not stationary, they may all have different expectations), rather μ is the analogous expectation with respect to the invariant distribution.

4.6 Reversibility

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

$$\iint f(x, y) \eta(dx) K(x, dy) = \iint f(y, x) \eta(dx) K(x, dy) \tag{14}$$

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 (14), 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 $\pi = \pi P$.

4.7 Total Variation Norm

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

$$\|\lambda\| = \sup_{A \in \mathcal{A}} \lambda(A) - \inf_{A \in \mathcal{A}} \lambda(A) \tag{15}$$

Clearly, we have

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

and hence

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

Conversely,

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

so

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

In summary,

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

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 (15).

4.8 Geometric Ergodicity

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

$$\sum_{n=1}^{\infty} r^n \|P^n(x, \cdot) - \pi(\cdot)\| < \infty, \qquad x \in \Omega. \tag{16}$$

(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

$$\|P^n(x, \cdot) - \pi(\cdot)\| \le M(x)s^n, \qquad x \in \Omega. \tag{17}$$

It is obvious that (16) implies (17), but the reverse implication is almost as obvious. If we assume (17), 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 (16) for any r such that $1 < r < 1/s$.

4.9 The Central Limit Theorem

The central limit theorem (CLT) for IID sequences of random variables says the following. Let X_1, X_2, ... be a sequence of IID random variables having expectation μ and standard deviation σ, and define $\overline{X}_n$ by (9), then

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{\mathcal{D}} \text{Normal}(0, \sigma^2) \tag{18}$$

this being known as the central limit theorem (in case $\sigma = 0$, the right hand side is interpreted as the degenerate normal distribution concentrated at zero).

We want to discuss the CLT for Markov chains, so again we have to go to functionals and again use the notation (11) and (12). Suppose X_1, X_2, ... is a geometrically ergodic Markov chain, π is its invariant distribution, and f is a real-valued function on the state space such that

$$E_\pi\{|f(X)|^{2+\varepsilon}\} = \int |f(x)|^{2+\varepsilon} \, \pi(dx) \tag{19}$$

exists for some $\varepsilon > 0$. Then

$$\sqrt{n}(\hat{\mu}_n - \mu) \xrightarrow{\mathcal{D}} \text{Normal}(0, \sigma^2), \tag{20}$$

where $\hat{\mu}_n$ and μ are given by (12) and (11), where

$$\sigma^2 = \gamma_0 + 2 \sum_{k=1}^{\infty} \gamma_k \tag{21}$$

and where

$$\gamma_n = \operatorname{cov}_\pi\{f(X_i), f(X_{i+n})\}, \tag{22}$$

the notation cov_π indicating covariances with respect to the stationary Markov chain having π as its initial distribution (Chan and Geyer, 1994, Theorem 2).

If we change assumptions for the theorem stated above slightly, adding reversibility to the assumptions and weakening (19) by only requiring it hold for $\varepsilon = 0$, then the conclusions still hold (Roberts and Rosenthal, 1997, Theorem 2.1, combined with a central limit theorem for rho-mixing stationary stochastic processes, Peligrad, 1986, Theorem 2.2, Remark 2.2, and Theorem 2.3, combined with the fact that any rho-mixing Markov chain is rho-mixing exponentially fast Bradley, 1986, Theorem 4.2). This theorem is, qualitatively, no worse than the CLT for IID. Both say the CLT holds for all functionals having second moments. Without reversibility we need a little bit more than second moments.

Geometric ergodicity is not necessary for a CLT. There are a lot of Markov chain CLT in the literature, but unlike the geometrically ergodic ones stated above, they do not provide any simple condition for which functionals (which f) have the CLT and which do not. They are thus not very usable in practice.

5 Monte Carlo

"Monte Carlo" is a cutesy name for the practice of learning about a probability distribution by simulating it. The term was coined in the 1950's when gambling was illegal almost everywhere (no legal gambling anywhere in the United States except in Las Vegas) and the casino at Monte Carlo was the most famous in the world. And gambling has something to do with randomness hence the term. It was a weak joke, now it is a colorless technical term designating a method in applied mathematics.

5.1 Ordinary Monte Carlo

Suppose there is a probability distribution π and there is an expectation μ given by (11) that is analytically intractable: we cannot calculate it either with pencil and paper methods or with a computer algebra system (like Mathematica or Maple).

The ordinary Monte Carlo (OMC) method says to simulate IID realizations X_1, X_2, ... from the distribution π and use $\hat{\mu}_n$ given by (12) as an

estimator (or Monte Carlo approximation) of (11).

Then the LLN (13) says our Monte Carlo approximation converges almost surely to the quantity we want to calculate as n goes to infinity, and the CLT says the difference between our Monte Carlo approximation and the quantity we want to calculate, which we call the Monte Carlo error, converges to a normal distribution at a root n rate.

So there is no mystery to OMC. It is just the most elementary of statistics: using the sample mean to estimate the population mean and using the law of large numbers and the central limit theorem for justification and error analysis. One difference is that since n is how many realizations we have the computer generate, we can always have n very large.

In order to avoid confusion when applying the Monte Carlo method to problems arising in statistics, we always emphasize that n is the *Monte Carlo sample size*, the number of simulations done by the computer, rather than anything else called "sample size" in the statistical problem being done.

OMC has only two drawbacks, one major and one minor. The major one is that it is very difficult to simulate IID realizations of any complicated multivariate distribution. Univariate distributions are easy to simulate. Devroye (1986) has hundreds of recipes that have appeared in the literature. They allow simulation of just about any univariate distribution that can be described. But there are almost no recipes for multivariate distributions: uniform on a box, uniform on a ball, and multivariate normal are the only multivariate distributions that are easy to simulate.

The minor drawback is that the "square root law" (the root n in the CLT) means that only limited precision is possible. To get 10 times the accuracy, one needs 100 times the Monte Carlo sample size. To get 100 times the accuracy, one needs 10,000 times the Monte Carlo sample size. At some point one just gives up. Arbitrary precision is not practical.

5.2 Markov Chain Monte Carlo

The Markov chain Monte Carlo (MCMC) method says to simulate an irreducible positive recurrent Markov chain $X_1, X_2, \ldots$ having π as its unique invariant distribution. We still use (12) as the estimator of (11).

We have now left the realm of elementary statistics. For justification we need a LLN and CLT for Markov chains, which are completely missing from many statistics programs.

It is important to emphasize that the LLN and CLT for geometrically ergodic Markov chains do not depend on the initial distribution (if the LLN holds for any initial distribution, then it holds for every initial distribution,

and similarly for the CLT Meyn and Tweedie, 2009, Proposition 17.1.6) because in practice we never use the stationary distribution as the initial distribution. If we knew how to generate even one sample from the stationary distribution, we could do that over and over and do OMC.

So in MCMC the samples X_1, X_2, ... are usually neither independent nor identically distributed and their distribution is not the distribution of interest. If the samples were independent, then they actually are IID and MCMC is in this case actually OMC. 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. So in practice MCMC provides a not independent, not identically distributed, sample from the distribution of interest.

MCMC has only two drawbacks, one major and one minor. The minor one is the same one that OMC has. MCMC obeys the square root law too, so only limited precision is practical. The major drawback is very different. MCMC easily simulates any multivariate distribution (Section 5.5 below), so it does not have the major drawback of OMC.

The major drawback of MCMC is that you are never quite sure that it has worked. This is a bit hard to explain, so let us consider a very special case. What is the probability that an OMC calculation estimates probability zero for an event A having true probability $\pi(A)$? This problem is solvable by intro statistics students: it is just the multiplication rule and the complement rule, and the answer is $[1 - \pi(A)]^n$. What is the answer to the same question for MCMC rather than OMC? Let T_A denote the hitting time for A (first time after time zero that the chain enters A). If the chain is geometrically ergodic, then there exist $r > 1$ such that

$$E_\pi\{r^{T_A}\} < \infty$$

(Nummelin, 1984, Proposition 5.19). Thus by Markov's inequality, there exists a constant $M < \infty$ such that

$$P_\pi(T_A \geq n) \leq M r^{-n}$$

which says the same thing as our answer for OMC except that we usually have no sharp bounds for M and r. With OMC we know that $M = 1$ and $r = 1/[1 - \pi(A)]$ will do. With MCMC we only know that some $M < \infty$ and $r > 1$ will do.

This is not of merely theoretical concern. In practical situations, it may take a very large number of iterations to get a sample that is reasonably representative of the invariant distribution, and there is usually no simple

calculation that tells us how many iterations are required. Theorems do exist that give bounds on how many iterations are required (Rosenthal, 1995; Łatuszyński and Niemiro, 2011; Łatuszyński, et al., 2013), but these bounds are very sloppy except in the simplest problems. In most practical MCMC applications, such bounds are useless if they can be computed at all.

In summary, OMC has the major drawback that you can't do it for complicated multivariate problems, and MCMC has the major drawback that you are never quite sure it has worked.

5.3 Unnormalized Probability Densities

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 MCMC samplers for probability models specified by unnormalized densities.

5.4 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 analytically intractable 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.

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). If one wants to do fixed scan and also wants reversible, one can use a so-called palindromic fixed scan (the same forwards and backwards), such as 1, 2, 3, 2, 1 for $d = 3$.

5.5 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

$$r(x, y) = \frac{h(y)q(y, x)}{h(x)q(x, y)} \tag{23}$$

and

$$a(x, y) = \min\bigl(1, r(x, y)\bigr). \tag{24}$$

(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 (23) 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*, (23) is called the *Hastings ratio*, (24) 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 (23) becomes

$$r(x, y) = \frac{h(y)}{h(x)} \tag{25}$$

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 4. *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

$$P(x, A) = m(x)I(x, A) + \int_A q(x, y)a(x, y)\, \mu(dy), \tag{26}$$

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

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

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

$$\iint f(x, y)h(x)q(x, y)a(x, y)\mu(dx)\mu(dy) \tag{27}$$

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

$$h(x)q(x, y)a(x, y) = h(y)q(y, x)a(y, x) \tag{28}$$

holding for all x and y except for those in a set making no contribution to (27) because the integrand is zero. Thus we may assume

$$h(x) > 0 \text{ and } q(x, y) > 0 \text{ and } a(x, y) > 0, \tag{29}$$

which implies $r(x, y) > 0$ and also implies (29) with x and y swapped. For (x, y) satisfying both (29) and (29) with x and y swapped, neither the numerator nor the denominator in (23) is equal to zero, and

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

The proof of the claim (28) 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. □

5.6 One Variable at a Time Metropolis-Hastings

A variant of the Metropolis-Hastings algorithm has elementary updates that update one variable at a time. These updates are then combined in a fixed or random scan like with the Gibbs sampler. We will not write out the details; see Geyer (2011, Section 1.12.5).

Nowadays we use the term "Metropolis-Hastings algorithm" to refer to the procedure described the preceding section and must use some long-winded term like that in the title of this section to refer to this algorithm. However, this is historically inaccurate. The original example in Metropolis, et al. (1953) was a sampler of the type described in this section not the type described in the preceding section.

The Gibbs sampler is a special case of the algorithm described in this section (Geyer, 2011, Section 1.12.6).

The algorithm described in this section is a special case of the algorithm described in the following section.

5.7 The Metropolis-Hastings-Green Algorithm

Many other MCMC algorithms have been put in the literature. They are all (as far as I know) special cases of the Metropolis-Hastings-Green algorithm (Green, 1995), which is just like the Metropolis-Hastings algorithm except that is allows proposals from distributions not defined by probability density functions and hence is inherently measure theoretic. We do not describe it here; see Geyer (2011, Section 1.17).

5.8 Harris Recurrence

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.

5.9 Variance Estimation

In order to estimate the accuracy of Monte Carlo approximations, we must estimate the asymptotic variance in the CLT (21). There are many methods of doing this (Geyer, 1992, Section 3; Geyer, 2011, Section 1.10). We will only discuss the simplest, which use the method of batch means.

A "batch" is a consecutive part of a time series such as a functional of a Markov chain. If $f(X_1)$, $f(X_2)$, ... is a functional of a Markov chain, then $f(X_{i+1})$, $f(X_{i+2})$, ..., $f(X_{i+b})$ is a *batch of length* b, and

$$\hat{\mu}_{ib} = \frac{1}{b} \sum_{j=1}^{b} f(X_{i+j})$$

is the corresponding *batch mean*. The Markov chain CLT says

$$\sqrt{b}(\hat{\mu}_{ib} - \mu) \xrightarrow{\mathcal{D}} \text{Normal}(0, \sigma^2)$$

and this holds regardless of i, so we expect

$$b(\hat{\mu}_{ib} - \hat{\mu}_n)^2 \tag{30}$$

to be a good estimate of σ^2 when $0 \ll b \ll n$, where $\ll$ means "a lot greater than." We need $0 \ll b$ in order for the CLT to hold at size b and we need the $b \ll n$ in order for the randomness in $\hat{\mu}_n$ to be much less than the randomness in $\hat{\mu}_{ib}$ so (30) is a good approximation to

$$b(\hat{\mu}_{ib} - \mu)^2.$$

The method of batch means has several subvarieties. The method of *overlapping batch means* uses all of the batches of length b.

$$\hat{\sigma}^2_{\text{olbm}} = \frac{b}{n - b + 1} \sum_{i=1}^{n-b+1} (\hat{\mu}_{ib} - \hat{\mu}_n)^2$$

The method of *nonoverlapping batch means* uses only nonoverlapping and abutting batches of length b.

$$\hat{\sigma}^2_{\text{nolbm}} = \frac{b}{\lfloor n/b \rfloor} \sum_{k=1}^{\lfloor n/b \rfloor} (\hat{\mu}_{(k-1)b+1,b} - \hat{\mu}_n)^2$$

The overlapping batch means estimator is somewhat more efficient (Meketon and Schmeiser, 1984), but not necessarily enough to be worth the extra computer time and storage (Geyer, 2011, Section 1.10, last paragraph). So we consider only the latter.

There is another distinction between the method of *consistent batch means* (CBM), which requires both b and n/b to go to infinity at a certain rate as n goes to infinity (Jones, et al., 2006), and the method of *inconsistent batch means* (IBM), which fixes the number of batches so the batch length is $\lfloor n/m \rfloor$ if there are m batches. It can then be shown (Geyer, 1992, Section 3.2) that the batches are asymptotically IID normal with mean μ and variance σ^2 so an ordinary t confidence interval gives a valid confidence interval for the quantity μ being approximated by MCMC. So the only penalty of IBM versus CBM is that one uses a t critical value rather than a z critical value in constructing the confidence interval. So long as the number of batches is moderately large (greater than 30), this penalty is negligible. CBM has the virtue that it is consistent (of course), so it can be easily used as a component of more complicated procedures (Jones, et al., 2006) but it requires stronger regularity conditions than the CLT itself. IBM has the virtue that it it valid whenever the CLT holds.

6 Drift Conditions

6.1 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

$$P^n(x, A) \ge \nu(A), \qquad A \in \mathcal{A} \text{ and } x \in C. \tag{31}$$

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

$$P_q(x, A) \ge \nu(A), \qquad A \in \mathcal{A} \text{ and } x \in C. \tag{32}$$

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).

6.2 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 (a Polish space is a complete separable metric space, and again $\mathbb{R}^d$ is an example).

A function f on a metric space is *lower semicontinuous* (LSC) if

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

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

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

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 5. *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

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

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

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

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 6. *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 5.5 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,

$$T(x, A) = \int_A q(x, y)a(x, y)\, dy, \tag{33}$$

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 (33) 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. □

6.3 Periodicity

Suppose C is a small set satisfying (31) 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 $A_0, \ldots, A_{d-1}$ 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

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

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 7. *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 (26), 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 8. *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

$$\Pr\bigl(X_{n+1} \in A_k \mid X_n \in A_k, h_{I_n}(X_n)\bigr) = \Pr\bigl(X_{n+1} \in A_k \mid h_{I_n}(X_n)\bigr) \tag{34}$$

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 (34) 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 $1, \ldots, k$. 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 $h_1(Y_0)$, $h_2(Y_1), \ldots, h_k(Y_{k-1})$ the two random elements $X_n = Y_0$ and $X_{n+1} = Y_k$ are conditionally independent. Hence

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

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. □

6.4 The Aperiodic Ergodic Theorem

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

Theorem 9. *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 10. *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 9 where λ is concentrated at the point x.

6.5 Drift Conditions in General

An abstract state space $(\Omega, \mathcal{A})$ where Ω is just a set having no other properties gives us little to work with in studying transience and recurrence. The chain is transient when it moves off to infinity (sort of), but on a bare set there is no direction toward infinity.

The idea of drift functions is to impose directions on a bare set. A drift function is just a nonnegative-valued function V on the state space; it may have extra restrictions, but different ones in different applications. Uphill on the drift function is toward infinity. Downhill on the drift function is toward the center.

Drift conditions work by comparing the functions V and PV. The latter is (by definition)

$$(PV)(x) = \int P(x, dy)V(y) = E\{V(X_{n+1}) \mid X_n = x\}.$$

If V is unbounded, the integral (conditional expectation) need not exist, in which case we say the value (in the loose sense) is $+\infty$, which always makes sense because V is nonnegative.

To simplify notation, we write $PV(x)$ instead of $(PV)(x)$. The former seems less clear, but it can only mean the latter, since we have no definition of a kernel multiplied on the right by a number (rather than by a function).

A notion that is useful in describing some drift functions is the following. 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).

6.6 The Drift Condition for Transience

Part of Theorem 8.0.2 in Meyn and Tweedie (2009) says the following. Suppose P is ψ-irreducible. Then a Markov chain with transition probability kernel P is transient if and only if there exists a bounded drift function V and a ψ-positive set A such that

$$PV(x) \ge V(x), \qquad x \in A^c$$

and the set

$$\left\{ x \in \Omega : V(x) > \sup_{y \in A} V(y) \right\}$$

is ψ-positive.

Meyn and Tweedie (2009) do not give any examples of using this drift condition, and I do not know of any. So we will not illustrate it here

6.7 The Drift Condition for Harris Recurrence

The following is Theorem 9.1.8 in Meyn and Tweedie (2009). (See Section 6.5 for the meaning of PV and the definition of unbounded off petite sets.)

Theorem 11. *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*

$$PV(x) \le V(x), \qquad x \notin C, \tag{35}$$

holds. Then the chain is Harris recurrent.

6.7.1 Example: AR(1) Time Series, Continued

Meyn and Tweedie (2009, Section 8.5.2) use this theorem to show that the AR(1) time series with $\rho = 1$ is Harris recurrent. Since their proof is rather complicated (it goes on for a page and a half), we won't try to duplicate it here.

Because of the symmetry of the normal distribution, if X_0, X_1, X_2, ... is an AR(1) Markov chain with $\rho = -1$ and $\sigma = s$ started at $X_0 = x$, then X_0, X_2, X_4, ... is an AR(1) Markov chain with $\rho = 1$ and $\sigma = s\sqrt{2}$ started at $X_0 = x$. Hence recurrence of the latter implies recurrence of the former.

This settles the case left open in Section 4.2.1 above. In case $\rho^2 = 1$, the AR(1) Markov chain is null recurrent.

6.8 The Drift Condition for Geometric Ergodicity

Recall the definition of "unbounded off petite sets" from Section 6.5. The following is part of Theorem 15.0.1 in Meyn and Tweedie (2009).

Theorem 12. *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 $V \geq 1$, and constants $b < \infty$ and $\lambda < 1$ such that*

$$PV(x) \leq \lambda V(x) + bI(x, C), \qquad x \in \Omega, \tag{36}$$

holds. Then the chain is geometrically ergodic.

The function V is referred to as a drift function and (36) as the drift condition for geometric ergodicity.

Theorem 12 has a near converse, which is another part of Theorem 15.0.1 in Meyn and Tweedie (2009).

Theorem 13. *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* (36) *holds. Moreover, there exist constants $r > 1$ and $R < \infty$ such that*

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

This shows that the function M in (17) can be taken to be a positive multiple of a drift function V. Taking expectations with respect to π of both sides of (36) 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 (17) 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 14. *The geometric drift condition* (36) *holds if and only if* V *is unbounded off petite sets and there exists a constant* $L < \infty$ *such that*

$$PV \le \lambda V + L. \tag{37}$$

6.8.1 Example: AR(1) Time Series, Continued

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 (37) with $\lambda = \rho^2$ and $L = 1 - \rho^2 + \sigma^2$.

6.8.2 Example: 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

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

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

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

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_t, λ_t), $t = 1, 2, \ldots$ (we use t rather than n for the time because we already have another n in this problem) as follows

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

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

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

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

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

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 $\varepsilon < 1/n$.

References

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

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

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. New York: Springer-Verlag.

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. (1992). Practical Markov chain Monte Carlo (with discussion). *Statistical Science*, **7**, 473–511.

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., pp. 3–48. Boca Raton, FL: Chapman & Hall/CRC.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

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

Hoel, P. G., Port, S. C., and Stone, C. J. (1972). *Introduction to Stochastic Processes*. Boston: Houghton Mifflin. Republished, Waveland Press, Prospect Heights, Illinois, 1986.

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

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

Jones, G. L., Haran, M., Caffo, B. S., and Neath, R. (2006). Fixed-width output analysis for Markov chain Monte Carlo. *Journal of the American Statistical Association*, **101**, 1537–1547.

Łatuszyński, K., Miasojedow, B., and Niemiro, W. (2013). Nonasymptotic bounds on the estimation error of MCMC algorithms. *Bernoulli*, **19**, 2033–2066.

Łatuszyński, K., and Niemiro, W. (2011). Rigorous confidence bounds for MCMC under a geometric drift condition. *Journal of Complexity*, **27**, 23–38.

Meketon, M. S. and Schmeiser, B. W. (1984). Overlapping batch means: Something for nothing? In *Proceedings of the 1984 Winter Simulation Conference*, edited by Sheppard, S., Pooch, U., and Pegden, D., pp. 227230. Piscataway, NJ: IEEE Press.

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.

Peligrad, M. (1986). Recent advances in the central limit theorem and its weak invariance principle for mixing sequences of random variables (a survey). In *Dependence in Probability and Statistics: A Survey of Recent Results* (Oberwolfach, 1985), edited by Eberlein, E., and Taqqu, M. S. Boston: Birkhäuser.

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

Rosenthal, J. S. (1995). Minorization conditions and convergence rates for Markov chain Monte Carlo. *Journal of the American Statistical Association*, **90**, 558–566.

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