3
Markov Chain Monte Carlo

This is a self-contained introduction to Markov Chain Monte Carlo (MCMC), a sampling method for estimating expectations which has revolutionized the practice of Bayesian inference. There are now many books on MCMC: I like Robert and Casella (2004) for its wider scope, although I'm now a bit behind the times. For insight, I recommend Besag et al. (1995) and Besag (2004). Grimmett and Stirzaker (2001, ch. 6) is an excellent one-chapter summary of Markov chains, consistent with that book's uniformly high standard of clarity and insight; I will give more detailed references to their ch. 6 below.

Warning! This is an introduction to the mathematics of MCMC. The practice of MCMC is a large and rapidly-evolving subject. Brooks et al. (2011) would be a good place to start.

Reminder: properties of expectation.

1. Monotonicity: if $X \geq 0$, then $\mathrm{E}(X) \geq 0$.
2. Linearity: $\mathrm{E}(aX + bY) = a\,\mathrm{E}(X) + b\,\mathrm{E}(Y)$.
3. Convexity: $\min \mathcal{X} \leq \mathrm{E}(X) \leq \max \mathcal{X}$.
4. Triangle inequality: $|\mathrm{E}(X)| \leq \mathrm{E}\{|X|\}$.
5. Indicator property: $\mathrm{E}(\mathbb{1}_{X=x} \mid A) = \Pr(X = x \mid A)$.
6. Law of Total Probability (LTP): if $\{B_i\}$ is a partition, then
$$\Pr(A) = \textstyle\sum_i \Pr(A \mid B_i) \cdot \Pr(B_i).$$
7. Law of Iterated Expectation (LIE):
$$\mathrm{E}[g(X,Y)] = \mathrm{E}[\mathrm{E}\{g(X,Y) \mid X\}].$$
8. Taking Out What is Known (TOWK):
$$\mathrm{E}\{g(X) \cdot h(X,Y) \mid X\} = g(X) \cdot \mathrm{E}\{h(X,Y) \mid X\}.$$
9. Double Whammy:
$$\mathrm{E}[g(X) \cdot h(X,Y)] = \mathrm{E}[g(X) \cdot \mathrm{E}\{h(X,Y) \mid X\}].$$
10. Markov's inequality: if $X \geq 0$, then $\Pr(X \geq a) \leq \mathrm{E}(X)/a$.

3.1 Markov chains

Let $\boldsymbol{X} := (X_0, X_1, \dots)$ be a sequence of random quantities with common realm $\mathcal{X}$, where $|\mathcal{X}| = r$, for some finite r. For Markov chains, it is common to refer to $\mathcal{X}$ as the *state-space* of X. I will write

$$\mathcal{X} := \{1, \dots, r\}$$

without loss of generality. $\boldsymbol{X}$ is a *Markov chain* exactly when

$$X_{t+1} \perp\!\!\!\perp \boldsymbol{X}_{0:(t-1)} \mid X_t \quad \text{for all } t = 0, 1, \dots, \tag{3.1}$$

where $\boldsymbol{X}_{i:j} := (X_i, \dots, X_j)$ for $i \leq j$ and $\emptyset$ otherwise. That is, X_{t+1} is conditionally independent of the 'past' given the 'present'. Using the equivalent representations of conditional independence given in Section 2.4, $\boldsymbol{X}$ is a Markov chain if and only if

$$\Pr(X_{t+1} = j \mid X_0 = x_0, \dots, X_t = i) = \Pr(X_{t+1} = j \mid X_t = i). \tag{3.2}$$

As a DAG,

$$X_0 \longrightarrow X_1 \longrightarrow X_2 \longrightarrow \cdots \tag{3.3}$$

In the case where these probabilities are invariant to t, the Markov chain is termed *homogeneous*. This is the case we are interested in. As $|\mathcal{X}|$ is finite, the transition probabilities of a homogeneous Markov chain can be packaged into an $r \times r$ matrix

$$P := \begin{pmatrix} p_{11} & \cdots & p_{1r} \\ \vdots & \ddots & \vdots \\ p_{r1} & \cdots & p_{rr} \end{pmatrix} \quad \text{where } p_{ij} := \Pr(X_{t+1} = j \mid X_t = i),$$

the same for all t. In other words, p_{ij} is the probability of going from i to j in one step. By construction, $p_{ij} \geq 0$ and $P\mathbf{1} = \mathbf{1}$, where $\mathbf{1}$ is an r-vector of ones. These are the defining properties of a *stochastic matrix*, although I will use the term 'transition matrix', which is common for Markov chains.

Let $p_{ij}(1) := p_{ij}$, and let $p_{ij}(n)$ represent the probability of going from i to j in exactly n steps. Then

$$\begin{aligned} p_{ij}(2) &= \mathrm{p}(X_2 = j \mid X_0 = i) \\ &= \textstyle\sum_k \Pr(X_2 = j \mid X_1 = k, X_0 = i) \cdot \Pr(X_1 = k \mid X_0 = i) \\ &= \textstyle\sum_k \Pr(X_2 = j \mid X_1 = k) \cdot \Pr(X_1 = k \mid X_0 = i) \qquad \text{by (3.1)} \\ &= \textstyle\sum_k p_{kj} \cdot p_{ik} = \sum_k p_{ik} \cdot p_{kj} = [P^2]_{ij}. \end{aligned}$$

By iterating this result we get $p_{ij}(n) = [P^n]_{ij}$. Now let $\mu_i(t) := \Pr(X_t = i)$, and $\boldsymbol{\mu}(t) := \big(\mu_1(t), \dots, \mu_r(t)\big)$. $\boldsymbol{\mu}(t)$ is a point in the $(r-1)$-dimensional simplex,

$$\mathbb{S}^{r-1} := \left\{ x \in \mathbb{R}^r : x_i \geq 0, \textstyle\sum_i x_i = 1 \right\}. \tag{3.4}$$

Then

$$\begin{aligned} \mu_j(n) &= \Pr(X_n = j) \\ &= \textstyle\sum_i \Pr(X_n = j \mid X_0 = i) \cdot \Pr(X_0 = i) \qquad \text{by the LTP} \\ &= \textstyle\sum_i p_{ij}(n) \cdot \mu_i(0), \end{aligned}$$

or, in matrix terms,

$$\boldsymbol{\mu}(n)^T = \boldsymbol{\mu}(0)^T P^n.$$

In other words, the n-step ahead probability distribution can be computed directly from the current probability distribution and the n-th power of the transition matrix.

Homogeneous Markov chains have many fascinating properties, but I am going to be very selective. P is termed *irreducible* if it is possible to get from every i to every j in a finite number of steps; i.e. if for every i, j there is an n for which $p_{ij}(n) > 0$ (this n possibly depending on i, j). In P, state i is termed aperiodic if $X_t = i$ can happen at irregular times. P is termed *aperiodic* if all of its states are aperiodic.[1] If $p_{ij} > 0$ then P is necessarily irreducible and aperiodic. The following is a standard result from Markov chain theory; see Grimmett and Stirzaker (2001, sec. 6.6) or Whittle (2000, ch. 9).

[1] In an irreducible P, the existence of a single aperiodic state implies that P is aperiodic.

Theorem 3.1. *Let P be the transition matrix for a homogeneous Markov chain with a finite state-space. If P is irreducible and aperiodic then there exists a unique probability vector $\boldsymbol{\pi}$ satisfying $\boldsymbol{\pi}^T P = \boldsymbol{\pi}^T$.*

These conditions are required for uniqueness, but not for existence. Brouwer's Fixed Point Theorem can be used to show that every P has a $\boldsymbol{\pi}$ satisfying the above property.[2] For $\boldsymbol{\pi} \mapsto P^T\boldsymbol{\pi}$ is a continuous map from $\mathbb{S}^{r-1}$ to $\mathbb{S}^{r-1}$, and therefore it must have a fixed point satisfying $\boldsymbol{\pi} = P^T\boldsymbol{\pi}$, or $\boldsymbol{\pi}^T = \boldsymbol{\pi}^T P$.

[2] Brouwer's FPT states that if $f : \mathcal{X} \to \mathcal{X}$ is continuous and $\mathcal{X}$ is a convex and compact subset of Euclidean space, then f has a fixed point x_0 satisfying $x_0 = f(x_0)$. It is trivial to prove when $\mathcal{X} \subset \mathbb{R}$ (a diagram suffices), but much harder to prove for $\mathcal{X} \subset \mathbb{R}^r$.

The probability vector $\boldsymbol{\pi}$ is termed the *stationary distribution* of P, because

$$\boldsymbol{\pi}^T P^n = \boldsymbol{\pi}^T P P^{n-1} = \boldsymbol{\pi}^T P^{n-1} = \cdots = \boldsymbol{\pi}^T,$$

with $P^0 = I$. In other words, if the distribution of X_t is equal to $\boldsymbol{\pi}$, then the distribution of X_{t+n} is equal to $\boldsymbol{\pi}$ for all $n > 0$.

3.2 *Convergence of Cesàro averages*

If $\boldsymbol{X}$ is a sequence of random quantities with a common state-space $\mathcal{X}$, and $g : \mathcal{X} \to \mathbb{R}$ is a specified function, the *Cesàro average* of $\{g(X_t); t \geq 1\}$ is defined as

$$\frac{1}{n}\sum_{t=1}^{n} g(X_t). \tag{3.5}$$

The crucial question for MCMC applications concerns the large-n behaviour of the Cesàro averages when $\boldsymbol{X}$ is a homogeneous Markov chain. In particular, do they converge as $n \to \infty$ and, if so, what do they converge to?[3]

[3] I am reliably informed that 'Chezaro' is the right pronounciation.

This question requires us to be precise about the meaning of 'convergence' when dealing with random quantities. There are several different types, but I will focus on convergence in mean square. Here is the general definition; if the state-space is finite the first condition is always satisfied.

Definition 3.1 (Convergence in mean square)**.** A sequence $X_1, X_2, \ldots$ converges in mean square to X exactly when $\mathrm{E}(|X_n|^2) < \infty$ for all n

and

$$\lim_{n\to\infty} \mathrm{E}\left\{(X_n - X)^2\right\} = 0.$$

This is written $X_n \xrightarrow{\text{m.s.}} X$.

Convergence in mean square is a strong form of convergence: it implies convergence in probability.[4] See Grimmett and Stirzaker (2001, ch. 7) for more details about different types of convergence.

[4] Easily proved using Markov's inequality.

In this section I provide a self-contained proof for a fundamental property of Markov chains, which is that if $X_0, X_1, \ldots$ is a homogeneous Markov chain with transition matrix P, and if P is irreducible and aperiodic, then for any initial distribution $\boldsymbol{\pi}_0$,

$$\frac{1}{n}\sum_{t=1}^{n} g(X_t) \xrightarrow{\text{m.s.}} \mathrm{E}\{g(X); \boldsymbol{\pi}\}$$

where $\mathrm{E}\{\cdot\,; \boldsymbol{\pi}\}$ is the expectation under the stationary distribution $X \sim \boldsymbol{\pi}$.[5] This means that if $x_0, x_1, \ldots, x_n$ is a realisation of $\boldsymbol{X}_{0:n}$, then the Cesàro average

$$\frac{1}{n}\sum_{t=1}^{n} g(x_t)$$

is a good estimate of $\mathrm{E}\{g(X); \boldsymbol{\pi}\}$ when n is sufficiently large. To anticipate the next section, if we can simulate a Markov chain with our target distribution as its stationary distribution, then we can use the Cesàro averages of that simulation to approximate any expectations of interest.

[5] In the case where the state-space is non-finite, this result would only hold if the expectation was well-defined; e.g. if g was a bounded continuous function.

As usual, I will concentrate on the case where the state-space is finite. There are three key results, of which this is the first.

Theorem 3.2. *Let P be the transition matrix for a homogeneous Markov chain with a finite state-space. If P is irreducible and aperiodic with stationary distribution $\boldsymbol{\pi}$ then there exists a $\lambda \in (0,1)$ and a positive constant c for which*

$$|p_{ij}(n) - \pi_j| \leq \lambda^n \cdot c \tag{3.6}$$

for all i, j; if all elements of P are positive, then $c = 1$.

Proof. This proof has two parts. First, I prove the special case where all the elements of P are strictly positive; then I generalize to the case where some of the elements of P may be zero.[6]

[6] I wish I could claim some credit for this beautiful proof, but it was sketched for me by Prof. Balint Toth. My original proof was much clunkier and used Perron's theorem.

So suppose that $p_{ij} > 0$. Define $\Pi := \mathbf{1}\boldsymbol{\pi}^T$, and note that

$$P\Pi = \Pi P = \Pi^2 = \Pi. \tag{†}$$

Because all elements of P are positive, there is an $\alpha \in (0,1)$ for which $P - \alpha\Pi \geq \mathbf{0}$, where the relation $A \geq \mathbf{0}$ indicates that $a_{ij} \geq 0$ for all i, j. Now define

$$Q := \frac{1}{1-\alpha}(P - \alpha\Pi),$$

and note that $Q \geq \mathbf{0}$ and $Q\mathbf{1} = 1$, so that Q is a transition matrix. Also note that

$$Q\Pi = \Pi Q = \Pi. \tag{‡}$$

Rearrange to give $P = (1-\alpha)Q + \alpha\Pi$, from which

$$\begin{aligned} P^n &= \left[(1-\alpha)Q + \alpha\Pi\right]^n \\ &= (1-\alpha)^n Q^n + \sum_{i=1}^{n} \binom{n}{i} (1-\alpha)^{n-i}\alpha^i \cdot \Pi \\ &= (1-\alpha)^n Q^n + \left(1 - (1-\alpha)^n\right)\Pi. \end{aligned}$$

In the second line the n-th power is expressed using the Binomial expansion, and all of the product terms in G and Π simplify to Π according to (†) and (‡). The third line recognises that because $\alpha \in (0,1)$ these are the probabilities of the Binomial(n,α) distribution. Finally, write $|A|$ for the matrix with entries $|a_{ij}|$. Then subtracting Π from both sides gives

$$|P^n - \Pi| = (1-\alpha)^n |Q^n - \Pi| \leq (1-\alpha)^n \, \mathbf{1}\mathbf{1}^T,$$

because Q^n and Π are both transition matrices. Taking the (i,j)-th element shows that $|p_{ij}(n) - \pi_j| \leq \lambda^n$, where $\lambda := (1-\alpha) \in (0,1)$. This proves that $c = 1$ in (3.6) in the special case where all of the elements of P are positive.

Now for the more general case. Because P is aperiodic, there is an n_0 for which all the elements of P^{n_0} are positive.[7] Let $\tilde{\alpha}$ and $\widetilde{Q}$ be the α and Q for P^{n_0}, following the same route as before. Write $P^n = P^{kn_0+m}$ where $k := \lfloor n/n_0 \rfloor$ and $m \in \{0, 1, \ldots, n_0 - 1\}$. Then (remembering that $\Pi P = \Pi$),

[7] If P is aperiodic then for each i, j there is an n_{ij} for which $p_{ij}(n) > 0$ for all $n \geq n_{ij}$. The n_0 in this proof is $n_0 := \max\{n_{ij}\}$.

$$\begin{aligned} P^n - \Pi &= \left[P^{n_0}\right]^k P^m - \Pi P^m \\ &= \left[(1-\tilde{\alpha})^k \widetilde{Q}^k + \left(1 - (1-\tilde{\alpha})^k\right)\Pi\right] P^m - \Pi P^m \\ &= (1-\tilde{\alpha})^k (\widetilde{Q}^k - \Pi) P^m, \end{aligned}$$

following the same route as before. So

$$\begin{aligned} |P^n - \Pi| &\leq (1-\tilde{\alpha})^k |\widetilde{Q}^k - \Pi| |P^m| \\ &\leq (1-\tilde{\alpha})^k \\ &= (1-\tilde{\alpha})^{\left\lfloor \frac{n}{n_0} \right\rfloor} \\ &\leq (1-\tilde{\alpha})^{\frac{n-1}{n_0}} \\ &= \left[(1-\tilde{\alpha})^{\frac{1}{n_0}}\right]^n \cdot c \end{aligned}$$

where $c := (1-\tilde{\alpha})^{-1/n_0} \geq 1$. This proves the result in (3.6), with $\lambda := (1-\tilde{\alpha})^{1/n_0} \in (0,1)$. □

The second stage is to prove that the proportion of time $\boldsymbol{X}$ spends in state j converges in mean square to the probability of the stationary distribution for state j.

I will use the *Big O notation* to simplify the proof; see Knuth (1973, sec. 1.2.11.1). If f and g are two functions with argument $n \in \mathbb{N}$, then we write $f(n) = \mathrm{O}\left(g(n)\right)$ exactly when there exists an n_0 and c such that $|f(n)| \leq c|g(n)|$ for all $n \geq n_0$. So $f(n) = \mathrm{O}(n)$ indicates that for n large enough, $|f(n)|$ is at most linear in n, and

$f(n) = \mathrm{O}(1)$ indicates that for n large enough, $f(n)$ is bounded; all constants are $\mathrm{O}(1)$. Some obvious properties, used below, are that $\mathrm{O}(n) + \mathrm{O}(n) = \mathrm{O}(n)$, $\mathrm{O}(1) \times \mathrm{O}(n) = \mathrm{O}(n)$, and $n^{-2}\,\mathrm{O}(n) = \mathrm{O}(n^{-1})$.

Theorem 3.3. *Under the conditions given in Theorem 3.2,*

$$\frac{1}{n}\sum_{t=1}^{n} \mathbb{1}_{X_t=j} \xrightarrow{m.s.} \pi_j,$$

for all j.

Proof. Fix j. Then

$$\begin{aligned}
&\left(\frac{1}{n}\sum_{t=1}^{n} \mathbb{1}_{X_t=j} - \pi_j\right)^2 \\
&\quad = \left(\frac{1}{n}\sum_{t=1}^{n} (\mathbb{1}_{X_t=j} - \pi_j)\right)^2 \\
&\quad = \frac{1}{n^2}\left(\sum_{t=1}^{n} (\mathbb{1}_{X_t=j} - \pi_j)^2 + 2\sum_{t=1}^{n}\sum_{s=t+1}^{n} (\mathbb{1}_{X_t=j} - \pi_j)\cdot(\mathbb{1}_{X_s=j} - \pi_j)\right),
\end{aligned}$$

after multiplying out. Hence

$$\begin{aligned}
&\mathrm{E}\left\{\left(\frac{1}{n}\sum_{t=1}^{n} \mathbb{1}_{X_t=j} - \pi_j\right)^2\right\} \\
&= \frac{1}{n^2}\underbrace{\sum_{t=1}^{n} \mathrm{E}\left\{(\mathbb{1}_{X_t=j} - \pi_j)^2\right\}}_{=:\,v_1(n)} + \frac{2}{n^2}\underbrace{\sum_{t=1}^{n}\sum_{s=t+1}^{n} \mathrm{E}\left\{(\mathbb{1}_{X_t=j} - \pi_j)\cdot(\mathbb{1}_{X_s=j} - \pi_j)\right\}}_{=:\,v_2(n)},
\end{aligned}$$

by Linearity. We need to show that the value of the expectation goes to zero as $n \to \infty$. As the value is bounded above by

$$\frac{1}{n^2}\Big(v_1(n) + 2|v_2(n)|\Big),$$

the rest of the proof consists in showing that $v_1(n) = \mathrm{O}(n)$ and $|v_2(n)| = \mathrm{O}(n)$. For then the value of the expectation is bounded above by

$$\frac{1}{n^2}\big(\mathrm{O}(n) + \mathrm{O}(n)\big) = \frac{1}{n^2}\,\mathrm{O}(n) = \mathrm{O}\left(\frac{1}{n}\right),$$

which goes to zero as $n \to \infty$, as required.

For $v_1(n)$, $\mathrm{E}\{(\mathbb{1}_{X_t=j} - \pi_j)^2\} = \mathrm{O}(1)$ by Convexity, and therefore the sum of n of these expectations is $\mathrm{O}(n)$, as required.

The behaviour of $v_2(n)$ is more interesting. For an individual term,

$$\begin{aligned}
&\mathrm{E}\left\{(\mathbb{1}_{X_t=j} - \pi_j)\cdot(\mathbb{1}_{X_s=j} - \pi_j)\right\} \\
&\quad = \mathrm{E}\left\{(\mathbb{1}_{X_t=j} - \pi_j)\cdot\mathbb{E}(\mathbb{1}_{X_s=j} - \pi_j \mid X_t)\right\} \quad \text{Double Whammy} \\
&\quad = \mathrm{E}\left\{(\mathbb{1}_{X_t=j} - \pi_j)\cdot(p_{X_t,j}(s-t) - \pi_j)\right\}, \qquad (\dagger)
\end{aligned}$$

where the last line uses

$$\begin{aligned}
\mathrm{E}(\mathbb{1}_{X_s=j} - \pi_j \mid X_t = i) &= \mathrm{E}(\mathbb{1}_{X_s=j} \mid X_t = i) - \pi_j \quad \text{Linearity} \\
&= \Pr(X_s = j \mid X_t = i) - \pi_j \quad \text{Indicator property} \\
&= p_{ij}(s-t) - \pi_j.
\end{aligned}$$

Hence

$$\begin{aligned}
|v_2(n)| &\le \textstyle\sum_t \sum_{s>t} |\mathrm{E}\left\{(\mathbb{1}_{X_t=j} - \pi_j)\cdot(\mathbb{1}_{X_s=j} - \pi_j)\right\}| \\
&= \textstyle\sum_t \sum_{s>t} |\mathrm{E}\left\{(\mathbb{1}_{X_t=j} - \pi_j)\cdot(p_{X_t,j}(i-t) - \pi_j)\right\}| && \text{from (†)} \\
&\le \textstyle\sum_t \sum_{s>t} \mathrm{E}\left\{|\mathbb{1}_{X_t=j} - \pi_j|\cdot|p_{X_t,j}(i-t) - \pi_j)|\right\} && \text{Triangle inequality} \\
&\le \textstyle\sum_t \sum_{s>t} \mathrm{E}\left\{|\mathbb{1}_{X_t=j} - \pi_j|\cdot\lambda^{s-t}\cdot c\right\} && \text{Theorem 3.2 and Monotonicity} \\
&= \textstyle\sum_t \sum_{s>t} \mathrm{O}(1)\cdot\lambda^{s-t},
\end{aligned}$$

the last step because $c \cdot \mathrm{E}\left\{|\mathbb{1}_{X_t=j} - \pi_j|\right\} = \mathrm{O}(1)$. Now lay out a tableau in t and s to identify all of the λ^{s-t} terms in the double sum:

	$t=1$	$t=2$	$t=3$	$\dots$	$t=n$
$s=1$					
$s=2$	λ^1				
$s=3$	λ^2	λ^1			
$s=4$	λ^3	λ^2	λ^1		
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	
$s=n$	λ^{n-1}	λ^{n-2}	λ^{n-3}	$\dots$	

For each column, the sum is less than $1/(1-\lambda)$, a positive constant not depending on n. There are n columns altogether, and thus

$$|v_2(n)| \le \textstyle\sum_t \sum_{s>t} \mathrm{O}(1)\cdot\lambda^{s-t} = \mathrm{O}(1)\cdot\mathrm{O}(n) = \mathrm{O}(n),$$

as needed to be shown. □

Theorem 3.3 supplies the required result for the final stage; happily this one is straightforward.

Theorem 3.4 (Mean square convergence of Cesàro averages)**.** *Under the conditions given in Theorem 3.2,*

$$\frac{1}{n}\sum_{t=1}^{n} g(X_t) \xrightarrow{m.s.} \mathrm{E}\{g(X);\boldsymbol{\pi}\},$$

for all g.

Proof. It is helpful to write

$$g(x) = \sum_{j=1}^{r} g_j\cdot\mathbb{1}_{x=j}, \quad \text{and} \quad \mathrm{E}\{g(X);\boldsymbol{\pi}\} = \sum_{j=1}^{r} g_j\cdot\pi_j, \qquad (\dagger)$$

writing g_j for $g(j)$. Then

$$\begin{aligned}
&\frac{1}{n}\textstyle\sum_t g(X_t) - \mathrm{E}\{g(X);\boldsymbol{\pi}\} \\
&\quad= \frac{1}{n}\textstyle\sum_t \left(g(X_t) - \mathrm{E}\{g(X);\boldsymbol{\pi}\}\right) \\
&\quad= \frac{1}{n}\textstyle\sum_t \sum_j g_j\cdot\left(\mathbb{1}_{X_t=j} - \pi_j\right) && \text{from (†)} \\
&\quad= \textstyle\sum_j g_j\cdot\frac{1}{n}\sum_t \left(\mathbb{1}_{X_t=j} - \pi_j\right) \\
&\quad= \textstyle\sum_j g_j\cdot\left(\frac{1}{n}\sum_t \mathbb{1}_{X_t=j} - \pi_j\right).
\end{aligned}$$

Squaring this and using the Cauchy-Schwarz Inequality gives

$$\left(\frac{1}{n}\sum_t g(X_t) - \mathrm{E}\{g(X);\boldsymbol{\pi}\}\right)^2 \leq \sum_j g_j^2 \cdot \sum_j \left(\frac{1}{n}\sum_t \mathbb{1}_{X_t=j} - \pi_j\right)^2.$$

The result then follows, on taking expectations, by Monotonicity, Linearity, and Theorem 3.3. □

There is one final point to make, so as not to give a misleading impression. Theorem 3.2 has geometric convergence to the stationary distribution, which is rapid. Generalizations of this result to a wider class of irreducible and aperiodic Markov chains (i.e. those with non-finite state-spaces) will not necessarily achieve the same rapid convergence.

3.3 *The Metropolis-Hastings (MH) algorithm*

Section 3.2 proved that Cesàro averages of irredicible and aperiodic Markov chains with finite state-spaces converge to the expectation under the stationary distribution. So our task, if we want to compute expectations with respect to some target distribution $\boldsymbol{\pi}$, is to construct an appropriate Markov chain with our target distribution as its stationary distribution. I will continue to assume that the state-space is finite.

First, restrict attention to a particular type of Markov chain. Let $\boldsymbol{\mu}$ be a probability distribution on $\mathcal{X}$. $\boldsymbol{\mu}$ satisfies *detailed balance* with respect to P exactly when

$$\mu_i P_{ij} = \mu_j P_{ji} \quad \text{for all } i, j \in 1, \ldots, r, \tag{3.7}$$

where $r = |\mathcal{X}|$. This is r^2 separate conditions, although r of these, when $i = j$, are automatically satisfied. These conditions are not hard to understand: with respect to $\boldsymbol{\mu}$, the probability that X_t is in i and then moves to j is the same as the probability that X_t is in j and then moves to i, for all i and j. For this reason, P is said to be *reversible* with respect to $\boldsymbol{\mu}$.

If $\boldsymbol{\mu}$ satisfies detailed balance with respect to P then it is straightforward to check that $\boldsymbol{\mu}$ is a stationary distribution for P:

$$\sum_i \mu_i P_{ij} = \sum_i \mu_j P_{ji} = \mu_j \sum_i P_{ji} = \mu_j,$$

i.e. $\boldsymbol{\mu}^T P = \boldsymbol{\mu}^T$. Grimmett and Stirzaker (2001, sec. 6.5) provide a very clear analogy to understand the relationship between the stationary distribution and reversibility.

If P is irredicible and aperiodic, ensuring that a stationary distribution exists and is unique, then a distribution which satisfies detailed balance with respect to P is the unique stationary distribution of P. So the objective is to take our target distribution $\boldsymbol{\pi}$ and construct an irreducible and aperiodic Markov chain with transition matrix P, for which $\boldsymbol{\pi}$ satisfies detailed balance with respect to P.

Surprisingly, not only is this possible, but it is easy, using the *Metropolis-Hastings (MH)* algorithm. The target distribution is $\boldsymbol{\pi}$.

The MH algorithm requires a *proposal distribution*

$$h(i \to j) := \Pr(\widetilde{X} = j \mid X_t = i), \tag{3.8a}$$

which is invariant to t; $\widetilde{X}$ is a 'temporary' random quantity with the same state-space as X_t. The MH algorithm for the transition matrix $P(i \to j) := P_{ij}$ is made up of two parts:

1. Letting $i = X_t$, sample $\widetilde{X} = j$ with probabilities $h(i \to j)$.
2. Set

$$X_{t+1} = \begin{cases} \widetilde{X} & \text{with probability } a(i \to j) \\ X_t & \text{with probability } 1 - a(i \to j), \end{cases} \tag{3.8b}$$

where

$$a(i \to j) := \min\left\{1, \frac{\pi_j \cdot h(j \to i)}{\pi_i \cdot h(i \to j)}\right\}. \tag{3.8c}$$

Eq. (3.8c) shows why the MH algorithm is so useful in Bayesian conditionalization. Bayesian statisticians would like to compute expectations with respect to the conditional distribution $\mathrm{p}(\theta, x \mid y^{\mathrm{obs}})$ i.e., this is their target distribution (see Chapter 1 and Chapter 2). Using the definition of conditional probability,

$$\mathrm{p}(\theta, x \mid y^{\mathrm{obs}}) = \frac{\mathrm{p}(\theta, x, y^{\mathrm{obs}})}{\mathrm{p}(y^{\mathrm{obs}})} \propto \mathrm{p}(\theta, x, y^{\mathrm{obs}}). \tag{3.9}$$

The dropped constant, $\mathrm{p}(y^{\mathrm{obs}})^{-1}$, is intractable, because computing it involves marginalizing over (Θ, X). Happily, in the MH algorithm the target distribution only enters as a ratio π_j / π_i, which means that all multiplicative constants in the target distribution cancel. *The modern revolution in Bayesian statistics is largely down to the fact that the MH algorithm side-steps the intractable normlizing constant in the conditional distribution.*

It is straightforward to see that the MH algorithm constructs a homogeneous Markov chain: the probability distribution for X_{t+1} depends only on the value of X_t, and not on the value of any previous X value, or on the value of t. Any sensible choice for h ought to ensure that $P(i \to j)$ is irreducible and aperiodic. So the only thing to be checked is the following.

Theorem 3.5 (Metropolis-Hastings)**.** *The target distribution $\boldsymbol{\pi}$ satisfies detailed balance with respect to the transition matrix implicitly defined in* (3.8).

Proof. The transition matrix P is defined implicitly, but it is straightforward to see that

$$P(i \to j) = \begin{cases} h(i \to j) \cdot a(i \to j) & j \neq i \\ 1 - \sum_{k \neq i} h(i \to k) \cdot a(i \to k) & j = i. \end{cases}$$

Now consider the detailed balance relation. We have to show that

$$\pi_i \cdot P(i \to j) = \pi_j \cdot P(j \to i) \quad \text{for all } i, j.$$

This automatically holds when $j = i$. In the case where $j \neq i$,

$$\begin{aligned}
\pi_i \cdot P(i \to j) &= \pi_i \cdot h(i \to j) \cdot a(i \to j) \\
&= \pi_i \cdot h(i \to j) \cdot \min\left\{1, \frac{\pi_j \cdot h(j \to i)}{\pi_i \cdot h(i \to j)}\right\} \\
&= \min\left\{\pi_i \cdot h(i \to j), \pi_j \cdot h(j \to i)\right\} \\
&= \pi_j \cdot h(j \to i) \cdot \min\left\{\frac{\pi_i \cdot h(i \to j)}{\pi_j \cdot h(j \to i)}, 1\right\} \\
&= \pi_j \cdot h(j \to i) \cdot a(j \to i) \\
&= \pi_j \cdot P(j \to i). \qquad \square
\end{aligned}$$

The nature of the MH algorithm is this: we get to choose the proposal distribution h at our convenience. We must choose something which is easy to simulate from and also for which is is easy to evaluate the probability $h(i \to j)$. Then we 'compensate' for our choice of h by a particular form for a to induce a transition matrix P. This form for a combines our proposal h and our target distribution $\boldsymbol{\pi}$ in such a way that $\boldsymbol{\pi}$ is the unique stationary distribution of P.

The great generality of the MH algorithm, which gives us almost a 'free choice' for h, is both its strength and its weakness. Among the uncountable number of possibilities for h, most will be very poor choices, in the sense that $\boldsymbol{\pi}(0), \boldsymbol{\pi}(1), \ldots$ will converge very slowly to the target $\boldsymbol{\pi}$. In the terms of Theorem 3.2, they will have $\lambda \approx 1$, and $c \gg 1$. But by a careful choice of h, we can try to construct a Markov chain with $\lambda \ll 1$ and $c \approx 1$.

Some common special cases of h have names. If h is symmetric, then the Metropolis-Hastings algorithm is simply the *Metropolis algorithm*. The advantage of a symmetric h is that $h(j \to i)/h(i \to j) = 1$, and so h never has to be evaluated in computing the acceptance probability. The Metropolis algorithm gives a simple insight into how the acceptance probability works. If the target density at the proposed value j is no lower than the currrent value i, then the proposal is always accepted. But if the target density is lower, then the proposal is accepted with probability π_j/π_i. So the chain always goes 'up-hill', but sometimes goes 'down-hill'.[8] One popular symmetric h is to make the distribution of the increment $|\widetilde{X} - X_t|$ symmetric about 0. This is the *random walk Metropolis algorithm*, discussed further in **??**.

[8] There is a close relationship between the Metropolis algorithm and optimization by *simulated annealing*; see Besag (2004).

If h is invariant to i, then the MH algorithm is an *independence sampler*. The proposals may be independent, but the evolution of the chain is not, because the acceptance of each proposal depends on the current state of the chain. This makes the independence sampler fundamentally different from simple Monte Carlo methods like *rejection sampling* and *importance sampling*.

* * *

There are some clever tunes we can play with the MH algorithm, which follow from the following basic result. The proof is straightforward and is omitted.

Theorem 3.6. *Let P and P' be transition matrices on the same state-space $\mathcal{X}$, and let $\alpha \in [0,1]$. Then*

$$A := \alpha P + (1-\alpha)P' \quad \textit{and} \quad B := PP'$$

are both transition matrices on $\mathcal{X}$. Furthermore, if $\boldsymbol{\pi}$ is a stationary distribution of both P and P', then it is a stationary distribution of both A and B.

This is a very useful result because the MH algorithm provides a simple way to construct lots of different transition matrices on the same state-space, all with the same unique stationary distribution. Transition matrix A represents a scheme in which P is chosen with probability α, and P' is chosen with probability $1-\alpha$. Transition matrix B represents a scheme in which first P is used, then P' is used. These results extend immediately to any finite set of transition matrices, provided that A is a convex combination in which the weights are non-negative and sum to one.

3.4 Gibbs sampling as a special case of MH

Section 3.3 noted that there are a lot of bad choices for the proposal distribution h. Gibbs sampling is a special case of the MH algorithm in which the choice for h is determined entirely by the target distribution $\boldsymbol{\pi}$, in such a way that the acceptance probability $a(i \to j)$ is always 1. Gibbs sampling applies whenever the random quantity of interest is a vector. It was in preparation for this development that I switched from P_{ij} to $P(i \to j)$ in the previous section.

Let the state space be $\boldsymbol{\mathcal{X}}$ and, for simplicity, let $\boldsymbol{\mathcal{X}} = \mathcal{X} \times \mathcal{X}'$, with just two elements. In keeping with the previous notation I will use i and j to index $\mathcal{X}$, and i' and j' to index $\mathcal{X}'$. In this notation, the target distribution is

$$\pi_{ii'} := \Pr(X = i, X' = i'),$$

and the transition matrix of the Markov chain is

$$P(ii' \to jj') := \Pr(X_{t+1} = j, X'_{t+1} = j' \mid X_t = i, X'_t = i').$$

In the MH algorithm, (3.8), the proposal distribution is now $h(ii' \to jj')$, and the acceptance probability is $a(ii' \to jj')$.

The Gibbs proposal updates one element of $\boldsymbol{X} = [X, X']$. This element is selected in some way (described further below); suppose that the first element is selected. The Gibbs proposal is defined using the target distribution $\boldsymbol{\pi}$, as

$$h(ii' \to jj') = \begin{cases} \Pr(X = j \mid X' = i'; \boldsymbol{\pi}) & j' = i' \\ 0 & \text{otherwise} \end{cases} \tag{3.10}$$

where

$$\Pr(X = j \mid X' = i'; \boldsymbol{\pi}) = \frac{\Pr(X = j, X' = i'; \boldsymbol{\pi})}{\Pr(X' = i'; \boldsymbol{\pi})} = \frac{\pi_{ji'}}{\sum_k \pi_{ki'}}.$$

In other words, using the target distribution, sample the proposed value for X conditional on the current value of X', and do not change X'. So the Gibbs sampler removes completely the 'free choice' of the proposal distribution h that is part of the general MH algorithm. Here is the key result.

Theorem 3.7 (Gibbs sampling)**.** *Under the Gibbs proposal in* (3.10)*, the MH algorithm in* (3.8) *has acceptance probability* 1*.*

Proof. As it is impossible for the proposed value of X' to be different from the current value, only the value of $a(ii' \to ji')$ matters. Then

$$\begin{aligned} a(ii' \to ji') &= \min\left\{1, \frac{\pi_{ji'}}{\pi_{ii'}} \cdot \frac{h(ji' \to ii')}{h(ii' \to ji')}\right\} \\ &= \min\left\{1, \frac{\pi_{ji'}}{\pi_{ii'}} \cdot \frac{\pi_{ii'} \big/ \sum_k \pi_{ki'}}{\pi_{ji'} \big/ \sum_k \pi_{ki'}}\right\} \\ &= \min\big\{1, 1\big\} = 1. \end{aligned}$$

□

Inspection of the Gibbs proposal and the proof of Theorem 3.7 shows that the result extends immediately to the case where $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2]$, a partition into two subsets with one or more elements each, and the whole of $\boldsymbol{X}_1$ is updated using the conditional distribution of $\boldsymbol{X}_1$ given $\boldsymbol{X}_2$ as the proposal, and leaving $\boldsymbol{X}_2$ unchanged.

One Gibbs proposal on its own cannot satisfy the convergence result in Theorem 3.4, because if $\boldsymbol{X}_2$ is never updated, then the Markov chain on $\boldsymbol{X}$ is not irreducible. But Theorem 3.6 shows that a set of Gibbs samplers which between them cover the whole of $\boldsymbol{X}$ will work: they will each target $\boldsymbol{\pi}$ and together they will be irreducible (barring pathological choices for $\boldsymbol{\pi}$).

As Theorem 3.6 suggests, there are two different schemes for a fixed partition of $\boldsymbol{X}$ into two or more groups, say $\boldsymbol{X} = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k]$. In scheme A, termed *random scan*, an element of the partition is chosen at random at each time-step, usually using uniform probabilities. In scheme B, termed *sequential scan*, the elements of the partition are updated sequentially. Sequential scan is the 'classic' Gibbs sampler; when statisticians refer to 'the Gibbs sampler', they usually mean Gibbs proposals applied sequentially to the elements of a partition of $\boldsymbol{X}$. Sequential scan is not homogeneous for the update of $\boldsymbol{X}_j$ alone, because the transition matrix depends on j, i.e. it is P_j for updating $\boldsymbol{X}_j$. But sequential scan *is* homogeneous taking the k updates as k mini-steps for the full transition matrix $P = P_k \cdots P_1$.

Here is one full step of Gibbs sampling under sequential scan,

starting with $x = (x_1, \ldots, x_m)$ and conditioning on $Y = y^{\text{obs}}$:

1. Sample $x'_1 \sim \mathrm{p}(x_1 \mid x_{2:m}, y^{\text{obs}})$
2. Sample $x'_2 \sim \mathrm{p}(x_2 \mid x'_1, x_{3:m}, y^{\text{obs}})$
3. Sample $x'_3 \sim \mathrm{p}(x_3 \mid x'_{1:2}, x_{4:m}, y^{\text{obs}})$

   $\vdots$

m. Sample $x'_m \sim \mathrm{p}(x_m \mid x'_{1:(m-1)}, y^{\text{obs}})$

to give the new value $x' = (x'_1, \ldots, x'_m)$. In this case the elements of X are processed singly, and there are m mini-steps: first X_1 is updated, then X_2, and so on. The distributions

$$\mathrm{p}(x_i \mid x_{-i}, y) = \mathrm{p}(x_i \mid x_{1:(i-1)}, x_{(i+1):m}, y)$$

are the 'full conditional distributions' of $\mathrm{p}(x, y)$, introduced in Section 2.7. As mentioned above, X can also be updated in blocks.

* * *

Suppose want to target $\mathrm{p}(\theta, x \mid y^{\text{obs}})$. If $\mathrm{p}(\theta, x, y)$ is available in symbolic form, it is possible to infer the symbolic forms of the full conditional distributions required for the Gibbs sampler. If these are familiar distributions, then the Gibbs proposals can be made using standard and highly efficient simulation algorithms. When we construct hierarchical models using 'off-the-shelf' distributions—Normal, Poisson, Gamma, Binomial, and so on—the full conditionals are often other off-the-shelf distributions. For unfamiliar distributions, a different type of proposal is needed. One possibility is a 1*D slice sampler*, see Neal (2003).

This approach has been exploited in several different software packages, typically based on the BUGS modelling language, which is used to express $\mathrm{p}(\theta, x, y)$ symbolically; see Lunn et al. (2009, 2013) for details. My preference is to use JAGS to do the simulation (Plummer, 2003, 2016). Software such as JAGS comes close to the holy grail of fully-automatic conditional sampling, given only the symbolic form of the joint distribution $\mathrm{p}(\theta, x, y)$ and the observations y^{obs}. Gibbs-sampling methods like JAGS work well for joint distributions expressed hierarchically. But they struggle with other kinds of joint distribution, such as those involving large spatial fields, which are often not expressed hierarchically (see, e.g., Cressie and Wikle, 2011).

3.5 Practical issues

The convergence of Cesàro averages theorem (Theorem 3.4) promises only asymptotic convergence of expectations (in mean square). We do not have an infinite amount of time and so we must plan for imperfect convergence. The two issues are: