7.2 Background material for Markov chains obtained by MCMC algorithms

In this section we survey a number of important convergence results for Markov chains constructed by MCMC algorithms. Since we have different uncountable state spaces for the chains in Section 7.1, we consider a framework with a general state space Ω. The theory for Markov chains on general state spaces is studied in great detail in Nummelin (1984) and Meyn & Tweedie (1993); see also Tierney (1994) and Roberts & Tweedie (2003) which are directed more to MCMC applications.

Consider a given probability distribution Π defined on Ω.† We can often construct a *time-homogeneous Markov chain* $Y_0, Y_1, \dots$ with state space Ω by some MCMC algorithm so that

$$P^m(x, F) \to \Pi(F) \tag{7.13}$$

for $F \subseteq \Omega$ and $x \in \Omega$, where

$$P^m(x, F) = P(Y_m \in F | Y_0 = x)$$

is the m-step *transition probability*. Time-homogeneity means that the *transition kernel*

$$P(x, F) = P(Y_{m+1} \in F | Y_m = x)$$

does not depend on $m \in \mathbb{N}_0$ (where $\mathbb{N}_0 = \{0, 1, 2, \dots\}$). The distribution of Y_0 is called the *initial distribution*. By time-homogeneity and the Markov property, the distribution of $(Y_0 \dots, Y_m)$ for any $m \in \mathbb{N}_0$ is determined by the initial distribution and the transition kernel.

Note that (7.13) implies that if $Y_0 \sim \Pi$ then $Y_m \sim \Pi$ for all $m \in \mathbb{N}_0$. Because, if e.g. Ω is discrete and $Y_0 \sim \Pi$, then for an arbitrary $y \in \Omega$,

$$\begin{aligned} P(Y_1 \in F) &= \sum_x P(x, F)\Pi(\{x\}) = \sum_x P(x, F) \lim_{m \to \infty} P^m(y, \{x\}) \\ &= \lim_{m \to \infty} \sum_x P(x, F) P^m(y, \{x\}) = \lim_{m \to \infty} P^{m+1}(y, F) = \Pi(F) \end{aligned}$$

and so by induction and time-homogeneity, $Y_m \sim \Pi$ for all $m \in \mathbb{N}$. MCMC algorithms are therefore naturally constructed so that Π is an *invariant distribution*, that is, if $Y_m \sim \Pi$ then $Y_{m+1} \sim \Pi$. Often the construction is so that *reversibility with respect to* Π holds, i.e. if $Y_m \sim \Pi$ then (Y_m, Y_{m+1}) and (Y_{m+1}, Y_m) are identically distributed, or equiva-

† For technical reasons we assume as in Nummelin (1984) and Meyn & Tweedie (1993) that Ω is equipped with a separable σ-algebra. This is satisfied for the space N_f and hence for the spaces E_n and E defined by (7.5) and (7.8), cf. Proposition B.1.

lently for all $F, G \subseteq \Omega$,

$$P(Y_m \in F, Y_{m+1} \in G, Y_m \neq Y_{m+1}) \\ = P(Y_m \in G, Y_{m+1} \in F, Y_m \neq Y_{m+1}) \quad \text{if } Y_m \sim \Pi. \tag{7.14}$$

This implies invariance, since if $Y_m \sim \Pi$,

$$P(Y_{m+1} \in F) = P(Y_m \in \Omega, Y_{m+1} \in F) = P(Y_m \in F, Y_{m+1} \in \Omega) \\ = \Pi(F).$$

7.2.1 Irreducibility and Harris recurrence

For $a > 0$ and $X \sim \Pi$, let $\mathcal{L}^a(\Pi)$ denote the class of functions $k : \Omega \to \mathbb{R}$ such that $\mathbb{E}|k(X)|^a < \infty$, and set $\mathcal{L}(\Pi) = \mathcal{L}^1(\Pi)$; if Π is specified by a normalised or unnormalised density h, we also write $\mathcal{L}^a(h)$ for $\mathcal{L}^a(\Pi)$, and $\mathcal{L}(h)$ for $\mathcal{L}(\Pi)$. Let $\Pi(k) = \mathbb{E}k(X)$ if $k \in \mathcal{L}(\Pi)$, and consider the *ergodic average* given by

$$\bar{k}_n = \frac{1}{n} \sum_{m=0}^{n-1} k(Y_m). \tag{7.15}$$

Below we establish two *strong laws of large numbers* (Propositions 7.3–7.4) for the estimate $\bar{k}_n$ of $\Pi(k)$ (such estimates are used several times in this book, see in particular Chapter 8). All what is needed is either irreducibility or Harris recurrence as defined in the sequel; convergence such as in (7.13) is a sufficient but not a necessary condition, cf. Section 7.2.2.

No matter whether an invariant distribution exists or not, we say that the chain is *irreducible* if there is a nonzero measure Ψ on Ω so that for any $x \in \Omega$ and $F \subseteq \Omega$ with $\Psi(F) > 0$, $P^m(x, F) > 0$ for some $m \in \mathbb{N}$; to stress the dependence on Ψ we also say that the chain is *Ψ-irreducible*. *Harris recurrence* means that the chain is Ψ-irreducible for some Ψ, and that for all $x \in \Omega$ and all $F \subseteq \Omega$ with $\Psi(F) > 0$,

$$P(Y_m \in F \text{ for some } m \,|\, Y_0 = x) = 1.$$

PROPOSITION 7.2 If an invariant distribution Π exists, Ψ-irreducibility implies the following properties:

(i) Π-irreducibility;

(ii) Π is the unique invariant distribution (up to null sets);

(iii) Π dominates Ψ, i.e. $\Pi(F) = 0$ implies $\Psi(F) = 0$;

(iv) there exists some $A \subset \Omega$ with $\Pi(A) = 0$ so that the chain restricted to $\Omega \setminus A$ is Harris recurrent and $\Omega \setminus A$ is absorbing, i.e. if $Y_0 \in \Omega \setminus A$ then $Y_m \in \Omega \setminus A$ for all $m \in \mathbb{N}_0$.

Proof. (i)–(iii) follow from Meyn & Tweedie (1993, Proposition 4.2.2 and Theorem 10.4.9). By Meyn & Tweedie (1993, Proposition 10.1.1), the chain is recurrent, that is, for any $x \in \Omega$ and $F \subseteq \Omega$ with $\Pi(F) > 0$, the mean number of times the chain visits F is infinite. Thus (iv) follows from Meyn & Tweedie (1993, Proposition 9.0.1). □

Obviously, (i) and (iii) are advantageous, since there can be much fewer sets $F \subseteq \Omega$ with $\Psi(F) > 0$ than $\Pi(F) > 0$ (this is exploited several times in Section 7.3). As requested, (ii) is telling us that if the chain has a limiting distribution, it must be Π. Finally, (iv) shows that irreducibility and Harris recurrence are essentially equivalent concepts; the disturbance is the nullset A.

PROPOSITION 7.3 If the chain is irreducible with invariant distribution Π and $k \in \mathcal{L}(\Pi)$, then there exists a set $A \subset \Omega$ so that $\Pi(A) = 0$ and with probability one, conditional on $Y_0 = x$ with $x \in \Omega \setminus A$, $\lim_{n\to\infty} \bar{k}_n = \Pi(k)$.

Proof. Combine (iv) in Proposition 7.2 with Proposition 7.4 below. □

To get rid of the nullset A in Proposition 7.3 Harris recurrence is needed:

PROPOSITION 7.4 The chain is Harris recurrent and has invariant distribution Π if and only if for all $k \in \mathcal{L}(\Pi)$, $\lim_{n\to\infty} \bar{k}_n = \Pi(k)$ with probability one and regardless of the initial distribution.

Proof. See Meyn & Tweedie (1993, Theorem 17.1.7). □

Proposition 7.4 includes the classical strong law of large numbers for i.i.d. random variables. To establish Harris recurrence, the following is useful. A set $C \subseteq \Omega$ is said to be *small* if there exist $m \in \mathbb{N}$ and a nonzero measure M on Ω such that

$$P^m(x, F) \geq M(F) \quad \text{for all } x \in C \text{ and } F \subseteq \Omega. \tag{7.16}$$

To stress the dependence on m and M we say that C is (m, M)-small. As we shall see later, small sets can be rather large; it may even occur that Ω is small.

PROPOSITION 7.5 † If the chain is Ψ-irreducible and there exist a small

† In Proposition 7.5 "small" may be replaced by the weaker concept of "petite", cf. Meyn and Tweedie 1993, but for simplicity we have excluded this in the formulation of the proposition. In fact, every small set is petite, and when the chain is irreducible and aperiodic, small and petite mean the same thing (Meyn & Tweedie 1993, Theorem 5.5.7).

set $C \subseteq \Omega$ and a function $V : \Omega \to [1, \infty)$ such that $\{x \in \Omega : V(x) \leq \alpha\}$ is small for all $\alpha > 1$ and

$$\mathbb{E}[V(Y_1)|Y_0 = x] \leq V(x) \quad \text{for all } x \in \Omega \setminus C, \tag{7.17}$$

then the chain is Harris recurrent.

Proof. See Meyn & Tweedie (1993, Theorem 9.1.8). □

We call (7.17) the *drift criterion for recurrence.*

7.2.2 *Aperiodicity and ergodicity*

We now consider when convergence results such as (7.13) are satisfied.

For an Ψ-irreducible chain, Ω can be partitioned into sets $D_0, \ldots, D_{d-1}$ and A so that $P(x, D_j) = 1$ for $x \in D_i$ and $j = i + 1 \bmod d$, and $\Psi(A) = 0$ (Meyn & Tweedie 1993, Theorem 5.4.4). If there is such a partition with $d > 1$ the chain is *periodic*, else it is *aperiodic*. We have aperiodicity if for example $P(x, \{x\}) > 0$ for some $x \in \Omega$. The following result is often useful.

PROPOSITION 7.6 An irreducible Markov chain with invariant distribution Π is aperiodic if and only if for some small C with $\Pi(C) > 0$ and some $n \in \mathbb{N}$, $P^m(x, C) > 0$ for all $x \in C$ and $m \geq n$.

Proof. See Proposition 5.3.4 in Roberts & Tweedie (2003) (since in the formulation of Proposition 7.6, "for some small" can be replaced by "for all small"). □

Clearly, periodicity implies that $P^m(x, D_0)$ cannot converge. In the sequel we therefore restrict attention to the aperiodic case.

For any initial distribution, let

$$Q^m(F) = \mathbb{E}P^m(Y_0, F) = P(Y_m \in F), \quad F \subseteq \Omega,$$

denote the marginal distribution of Y_m. In the special case where the initial distribution is concentrated at a given $x \in \Omega$, $Q^m(F) = P^m(x, F)$. To measure the distance between Q^m and Π we introduce the *total variation norm* given by

$$\|\mu - \nu\|_{\text{TV}} = \sup_{F \subseteq \Omega} |\mu(F) - \nu(F)|$$

for any two probability distributions μ and ν defined on Ω. Note that $\|\mu - \nu\|_{\text{TV}} \leq 1$, $\|\mu - \nu\|_{\text{TV}} = 0$ if $\mu = \nu$, and $\|\mu - \nu\|_{\text{TV}} = 1$ if μ and ν have disjoint supports. Note also that convergence of Q^m in total variation norm to Π is a rather strong form of convergence, which implies convergence of Q^m in distribution to Π.

PROPOSITION 7.7 Suppose the chain has invariant distribution Π. Then the following properties hold:

(i) Regardless of the initial distribution, $\|Q^m - \Pi\|_{\mathrm{TV}}$ is nonincreasing in m.

(ii) If the chain is irreducible and aperiodic, there exists $A \subset \Omega$ so that $\Pi(A) = 0$ and for all $x \in \Omega \setminus A$, $\lim_{m \to \infty} \|P^m(x, \cdot) - \Pi\|_{\mathrm{TV}} = 0$.

(iii) The chain is Harris recurrent and aperiodic if and only if $\lim_{m \to \infty} \|Q^m - \Pi\|_{\mathrm{TV}} = 0$ regardless of the initial distribution.

Proof. (i) follows from Meyn & Tweedie (1993, Theorem 13.3.2). Combining (iv) in Proposition 7.3 with Meyn & Tweedie (1993, Theorem 13.3.3) we obtain (ii)–(iii). □

We say that the chain is *ergodic* if it is both Harris recurrent and aperiodic. Particularly, ergodicity implies (7.13) regardless of the initial state.

7.2.3 Geometric and uniform ergodicity

We discuss now the rate of convergence, and consider finally a central limit theorem.

For $x \in \Omega$ and functions $V : \Omega \to [1, \infty)$ with $\Pi(V) < \infty$, define the V-*norm* of $P^m(x, \cdot) - \Pi$ by

$$\|P^m(x, \cdot) - \Pi\|_V = \frac{1}{2} \sup_{|k| \leq V} \left| \mathbb{E}[k(Y_m) | Y_0 = x] - \Pi(k) \right|$$

where the supremum is over all functions $k : \Omega \to \mathbb{R}$ with $|k(\cdot)| \leq V(\cdot)$. Note that $\|P^m(x, \cdot) - \Pi\|_V = \|P^m(x, \cdot) - \Pi\|_{\mathrm{TV}}$ if $V = 1$, and $\|P^m(x, \cdot) - \Pi\|_{V_1} \leq \|P^m(x, \cdot) - \Pi\|_{V_2}$ if $V_1 \leq V_2$. Following Meyn & Tweedie (1993) the chain is said to be V-*geometrically ergodic* if it is Harris recurrent with invariant distribution Π and there exists a constant $r > 1$ such that for all $x \in \Omega$,

$$\sum_{m=1}^{\infty} r^m \|P^m(x, \cdot) - \Pi\|_V < \infty. \tag{7.18}$$

We say that the chain is *geometrically ergodic* if it is V-geometrically ergodic for some V. In fact (7.18) is equivalent to that

$$\|P^m(x, \cdot) - \Pi\|_V \leq A(x) \tilde{r}^m \tag{7.19}$$

where $A(x) < \infty$ and $\tilde{r} < 1$ (Meyn & Tweedie 1993, p. 355). This shows both that the rate of convergence is geometric and that the chain is ergodic.

When we have geometric ergodicity with $A(\cdot) \propto V(\cdot)$ in (7.19), the chain is said to be *V-uniformly ergodic*. This is equivalent to that

$$\lim_{m\to\infty} \sup_{x\in\Omega} \|P^m(x,\cdot) - \Pi\|_V = 0,$$

cf. Meyn & Tweedie (1993, Theorem 16.2.1). If we have geometric ergodicity for a constant function A in (7.19), the chain is said to be *uniformly ergodic*. This is then equivalent to that

$$\lim_{m\to\infty} \sup_{x\in\Omega} \|P^m(x,\cdot) - \Pi\|_{\text{TV}} = 0,$$

but even more than that is known:

PROPOSITION 7.8 Suppose the chain has invariant distribution Π. Then uniform ergodicity is equivalent to that Ω is small. If it is (n, M)-small, then

$$\|P^m(x,\cdot) - \Pi\|_{\text{TV}} \le (1 - M(\Omega))^{m/n}, \quad m \in \mathbb{N}. \tag{7.20}$$

Proof. This follows from Meyn & Tweedie (1993, Theorem 16.2.2); see also Theorem 9.1.1 in Roberts & Tweedie (2003). □

To establish V-uniform ergodicity the following is useful.

PROPOSITION 7.9 Suppose the chain is aperiodic and irreducible with invariant distribution Π and there exist a function $V : \Omega \to [1, \infty)$, a small set $C \subseteq \Omega$, and constants $a < 1$ and $b < \infty$ such that for all $x \in \Omega$,

$$\mathbb{E}[V(Y_1)|Y_0 = x] \le aV(x) + b\mathbf{1}[x \in C]. \tag{7.21}$$

Then $\Pi(V) < \infty$, the chain is V-uniformly ergodic, and (7.18) can be sharpened to

$$\sum_{m=1}^{\infty} r^m \|P^m(x,\cdot) - \Pi\|_V \le \tilde{r}V(x) \tag{7.22}$$

for some constants $r > 1$ and $\tilde{r} < \infty$ and for all $x \in \Omega$.

Proof. This follows from Meyn & Tweedie (1993, Theorems 15.0.1 and 16.0.2), noticing that $\Pi(V) \le b/(1-a)$ is finite by (7.21), and that the chain is Harris recurrent by Meyn & Tweedie (1993, Lemma 15.2.8) and Proposition 7.5. □

We call (7.21) the *geometric drift condition*. Note that the bounds in (7.19) and (7.22) provide only qualitative results, since we do not know e.g. the value of $\tilde{r}$. Quantitative results for geometrically ergodic chains can be found in e.g. Rosenthal (1995), but as discussed in Møller (1999a) the results seem of very limited use for simulation of spatial point processes. This is also the case for uniformly ergodic chains: the upper bound given by (7.20) provides only a very rough estimate of the rate of convergence, cf. Example 7.2 and Remark 7.9.

Many algorithms for spatial point processes are V-uniformly ergodic but not uniformly ergodic, cf. Remark 7.9. Fortunately, geometric ergodicity (and hence V-uniform ergodicity) implies the following *central limit theorem.*

PROPOSITION 7.10 Suppose the chain is geometrically ergodic with invariant distribution Π, and k is a real function defined on Ω so that either

(i) $k \in \mathcal{L}^{2+\epsilon}(\Pi)$ for some $\epsilon > 0$, or

(ii) the chain is reversible and $k \in \mathcal{L}^2(\Pi)$, or

(iii) the chain is V-uniformly ergodic and $k^2 \leq V$.

Define

$$\bar{\sigma}^2 = \mathbb{V}\mathrm{ar}(k(Y_0)) + 2\sum_{m=1}^{\infty} \mathbb{C}\mathrm{ov}(k(Y_0), k(Y_m))$$

where the variance and covariances are calculated under the assumption that $Y_0 \sim \Pi$. Then $\bar{\sigma}^2$ is well-defined and finite, and regardless of the initial distribution, $\sqrt{n}(\bar{k}_n - \Pi(k))$ converges in distribution to $N(0, \bar{\sigma}^2)$ as $n \to \infty$ (in the special case $\bar{\sigma}^2 = 0$, convergence in distribution to $N(0,0)$ means that $\sqrt{n}(\bar{k}_n - \Pi(k))$ converges with probability one to the degenerate distribution concentrated at 0).

Proof. (i) See Chan & Geyer (1994, Theorem 1). (ii) See Roberts & Rosenthal (1997, Corollary 3). (iii) See Meyn & Tweedie (1993, Theorem 17.0.1). □

REMARK 7.8 Proposition 7.10 extends easily to the multivariate case. For example, in the bivariate case, for functions $k^{(1)}$ and $k^{(2)}$ which satisfy either (i), (ii), or (iii) in Proposition 7.10, $\sqrt{n}(\bar{k}_n^{(1)} - \Pi(k^{(1)}), \bar{k}_n^{(2)} - \Pi(k^{(2)}))$ is asymptotically normally distributed with mean $(0,0)$ and a covariance matrix with entries

$$\bar{\sigma}_{ij} = \mathbb{C}\mathrm{ov}(k^{(i)}(Y_0), k^{(j)}(Y_0)) + \sum_{m=1}^{\infty} \mathbb{C}\mathrm{ov}(k^{(i)}(Y_0), k^{(j)}(Y_m)) + \sum_{m=1}^{\infty} \mathbb{C}\mathrm{ov}(k^{(j)}(Y_0), k^{(i)}(Y_m))$$

for $i, j = 1, 2$, where the covariances are calculated under the assumption that $Y_0 \sim \Pi$.

7.3 Convergence properties of algorithms

In the following we investigate the convergence properties of the Markov chains generated by the Metropolis-Hastings algorithms in Section 7.1.

7.3.1 The conditional case

Consider the Markov chain generated by Algorithm 7.1, with unnormalised target density π given by (7.1) and state space E_n, assuming that $Y_0 \in E_n$, cf. Remark 7.1. For $\bar{x} \in E_n$ and $F \subseteq E_n$, the chain has transition kernel

$$\begin{aligned} P(\bar{x}, F) =& \frac{1}{n} \sum_{i=1}^{n} \int_B \mathbf{1}[(x_1, \ldots, x_{i-1}, \xi, x_{i+1}, \ldots, x_n) \in F] q_i(\bar{x}, \xi) \alpha_i(\bar{x}, \xi) \mathrm{d}\xi \\ &+ r(\bar{x}) \mathbf{1}[\bar{x} \in F] \end{aligned} \tag{7.23}$$

where the acceptance probability $\alpha_i(\bar{x}, \xi)$ is defined by (7.3) and where

$$r(\bar{x}) = \frac{1}{n} \sum_{i=1}^{n} \int_B \mathbf{1}[(x_1, \ldots, x_{i-1}, \xi, x_{i+1}, \ldots, x_n) \in F] q_i(\bar{x}, \xi)(1 - \alpha_i(\bar{x}, \xi)) \mathrm{d}\xi$$

is the probability of remaining at $\bar{x}$. In Proposition 7.11 below we refer to the chain as the Metropolis-Hastings chain, and in the proof we use only the part of (7.23) corresponding to accepted proposals. We also refer to the "proposal" chain which is obtained by always accepting proposals, i.e. it has transition kernel

$$Q(\bar{x}, F) = \frac{1}{n} \sum_{i=1}^{n} \int_B \mathbf{1}[(x_1, \ldots, x_{i-1}, \xi, x_{i+1}, \ldots, x_n) \in F] q_i(\bar{x}, \xi) \mathrm{d}\xi$$

for $\bar{x} \in E_n$ and $F \subseteq E_n$.

PROPOSITION 7.11 The following properties hold for the fixed number of points Metropolis-Hastings Algorithm 7.1.

(i) The Metropolis-Hastings chain is reversible with respect to π.

(ii) Suppose that $E_n = B^n$ and the proposal chain is Ψ-irreducible, and for all $\bar{x} = (x_1, \ldots, x_n) \in B^n$, $i \in \{1, \ldots, n\}$, and $\xi \in B$,

$$q_i(\bar{x}, \xi) > 0 \;\Rightarrow\; q_i(\bar{y}, x_i) > 0 \tag{7.24}$$

where $\bar{y} = (x_1, \ldots, x_{i-1}, \xi, x_{i+1}, \ldots, x_n)$. Then the Metropolis-Hastings chain is Ψ-irreducible.

(iii) Suppose that the Metropolis-Hastings chain is irreducible, and there exist $\epsilon > 0$, $x_2, \ldots, x_n \in B$, and $D \subseteq B$ such that $|D| > 0$ and for all $x_1, \xi \in D$, we have that $\pi(\bar{x}) > 0$, $\pi(\xi, x_2, \ldots, x_n) > 0$, and

$$\min\{q_1(\bar{x}, \xi), \pi(\xi, x_2, \ldots, x_n) q_1((\xi, x_2, \ldots, x_n), x_1)/\pi(\bar{x})\} \geq \epsilon \tag{7.25}$$