A short course in Markov chains (second part)

Randal Douc

Chapter 1

Ergodicity and the Law of Large Numbers

We now focus on properties of Markov chains with unique invariant probability measures. We will show that such Markov chains are ergodic in a sense linked to ergodic dynamical systems. This, in turn, allows us to apply the Birkhoff ergodic theorem and obtain a law of large numbers.

Let us start with a refresher on ergodic results.

1.1 Dynamical systems

DEFINITION 1.1 . (Dynamical system) A dynamical system $\mathscr{D}$ is a quadruple $\mathscr{D} = (\Omega, \mathscr{F}, \mathbb{P}, T)$ where $(\Omega, \mathscr{F}, \mathbb{P})$ is a probability space and $T : \Omega \to \Omega$ is a measurable mapping such that $\mathbb{P} = \mathbb{P} \circ T^{-1}$.

LEMMA 1.2 . The collection of sets $\mathscr{I} = \{A \in \mathscr{F} \ : \ \mathbf{1}_A = \mathbf{1}_A \circ T\}$ is a σ-field and any set in $\mathscr{I}$ is called an invariant set.

PROOF. Clearly, $\Omega \in \mathscr{I}$. Moreover, if $A \in \mathscr{I}$, then for all $\omega \in \Omega$,

$$\mathbf{1}_{A^c}(\omega) = 1 - \mathbf{1}_A(\omega) = 1 - \mathbf{1}_A \circ T(\omega) = \mathbf{1}_{A^c} \circ T(\omega),$$

showing that $A^c \in \mathscr{I}$. Now, consider a countable family $(A_i)_{i\in\mathbb{N}}$ of sets in $\mathscr{I}$. Then $\omega \in \cap_{i\in\mathbb{N}} A_i$ if and only if $\omega \in A_i$ for all $i \in \mathbb{N}$, which is in turn equivalent to $T(\omega) \in A_i$ for all $i \in \mathbb{N}$. Thus, we have shown that $\omega \in \cap_{i\in\mathbb{N}} A_i$ if and only if $T(\omega) \in \cap_{i\in\mathbb{N}} A_i$. This shows that $\mathscr{I}$ is a σ-field. ■

DEFINITION 1.3 . (Ergodicity) A dynamical system $(\Omega, \mathscr{F}, \mathbb{P}, T)$ is said to be ergodic if $\mathscr{I}$ is trivial, that is, $A \in \mathscr{I}$ implies that either $\mathbb{P}(A) = 0$ or $\mathbb{P}(A) = 1$.

Our main interest is that ergodic dynamical systems satisfy Birkhoff's ergodic theorem, as detailed below. We first define the iterates of T by $T^0 = \mathrm{Id}$, and for all $k \geqslant 1$, $T^k = T^{k-1} \circ T$.

THEOREM 1.4 . (Birkhoff's theorem) Let $\mathscr{D} = (\Omega, \mathscr{F}, \mathbb{P}, T)$ be an ergodic dynamical system and let $h \in \mathsf{L}_1(\Omega)$. Then,

$$\lim_{n\to\infty} n^{-1} \sum_{k=0}^{n-1} h \circ T^k = \mathbb{E}[h] \ , \quad \mathbb{P} - a.s.$$

1.2 Markov chains and ergodicity

Let us now relate Markov chains to dynamical systems. On the canonical path space, $X_k(\omega) = \omega_k$ and $\mathbb{P}_\nu$ denotes the law of the chain with initial distribution ν. Recall the shift operator $S : \mathsf{X}^{\mathbb{N}} \ni \omega \mapsto \omega' \in \mathsf{X}^{\mathbb{N}}$ where $\omega = (\omega_i)_{i\in\mathbb{N}}$ and $\omega' = (\omega_{i+1})_{i\in\mathbb{N}}$. It is important to note that in general $(\mathsf{X}^{\mathbb{N}}, \mathscr{X}^{\otimes\mathbb{N}}, \mathbb{P}_\nu, S)$ is not a dynamical system unless the initial distribution ν is invariant with respect to P.

LEMMA 1.5 . Let P be a Markov kernel admitting an invariant probability measure π. Then, the quadruple $(\mathsf{X}^{\mathbb{N}}, \mathscr{X}^{\otimes\mathbb{N}}, \mathbb{P}_\pi, S)$ is a dynamical system.

PROOF. Indeed, the relation $\mathbb{P}_\pi = \mathbb{P}_\pi \circ S^{-1}$ is equivalent to the fact that for any $A \in \mathscr{X}^{\otimes\mathbb{N}}$, $\mathbb{P}_\pi(A) = \mathbb{P}_\pi(X_{0:\infty} \in A) = \mathbb{P}_\pi\left(X_{0:\infty} \in S^{-1}(A)\right) = \mathbb{P}_\pi(X_{1:\infty} \in A)$, which is a consequence of the fact that the sequence of random variables $\{X_k \,:\, k \in \mathbb{N}\}$ is strongly stationary under $\mathbb{P}_\pi$. ∎

The relation $\mathbb{P}_\pi = \mathbb{P}_\pi \circ S^{-1}$ tells us that for any $A \in \mathscr{X}^{\otimes\mathbb{N}}$, we have

$$\mathbb{E}_\pi[\mathbf{1}_A] = \mathbb{P}_\pi(A) = \mathbb{P}_\pi \circ S^{-1}(A) = \mathbb{P}_\pi(S^{-1}(A)) = \mathbb{E}_\pi[\mathbf{1}_{S^{-1}(A)}] = \mathbb{E}_\pi[\mathbf{1}_A \circ S].$$

This, in turn, implies that for any non-negative measurable function h on $\mathsf{X}^{\mathbb{N}}$,

$$\mathbb{E}_\pi[h] = \mathbb{E}_\pi[h \circ S].$$

We now provide conditions under which this dynamical system is ergodic. In what follows, $\mathscr{F}_k = \sigma(X_{0:k})$. Since invariant sets A belong to $\mathscr{X}^{\otimes\mathbb{N}}$, it can be useful to approximate A by sets in $\mathscr{F}_k$ for a suitable choice of k.

LEMMA 1.6 . (The approximation lemma) Any set $A \in \mathscr{X}^{\otimes\mathbb{N}}$ satisfies the following approximation property:

- for all $\delta > 0$, there exist $k \in \mathbb{N}$ and $B \in \mathscr{F}_k$ such that

$$\mathbb{E}_\pi[|\mathbf{1}_A - \mathbf{1}_B|] \leqslant \delta. \tag{1.1}$$

PROOF. This is a typical use of the π-λ theorem. Consider the class $\mathscr{M}$ of sets $A \in \mathscr{X}^{\otimes\mathbb{N}}$, for which the approximation (1.1) holds.

- If $A_0, A_1 \in \mathscr{M}$ and $A_0 \subset A_1$, then $A_1 \setminus A_0 \in \mathscr{M}$. This is actually immediate from the following identities, valid for all sets A_0, A_1, B_0, B_1,

$$\mathbf{1}_{A_1\setminus A_0} - \mathbf{1}_{B_1\setminus B_0} = \mathbf{1}_{A_1}\mathbf{1}_{A_0^c} - \mathbf{1}_{B_1}\mathbf{1}_{B_0^c} = \mathbf{1}_{A_1}\left(\mathbf{1}_{A_0^c} - \mathbf{1}_{B_0^c}\right) + (\mathbf{1}_{A_1} - \mathbf{1}_{B_1})\,\mathbf{1}_{B_0^c},$$

  which implies $\mathbb{E}_\pi[|\mathbf{1}_{A_1\setminus A_0} - \mathbf{1}_{B_1\setminus B_0}|] \leqslant \mathbb{E}_\pi[|\mathbf{1}_{A_0} - \mathbf{1}_{B_0}|] + \mathbb{E}_\pi[|\mathbf{1}_{A_1} - \mathbf{1}_{B_1}|]$.
- If $A_n \uparrow A$ where $A_n \in \mathscr{M}$ and $A_n \subset A_{n+1}$ for all $n \geqslant 0$, then, setting $A = \cup_n A_n$, we have

$$\lim_{n\to\infty} \mathbf{1}_{A_n} = \mathbf{1}_A,$$

  and this immediately implies that $A \in \mathscr{M}$ (check it carefully).

Then, $\mathscr{M}$ is a Dynkin system that contains the algebra $\cup_{k=0}^{\infty}\mathscr{F}_k$. By the π-λ theorem, it contains $\sigma(\cup_{k=0}^{\infty}\mathscr{F}_k) = \mathscr{X}^{\otimes\mathbb{N}}$. This finishes the proof. ∎

We now have all the tools to state and prove an extremely powerful result on Markov chains and ergodicity.

THEOREM 1.7 . Let P be a Markov kernel on $\mathsf{X}\times\mathscr{X}$. Assume that P admits a unique invariant probability measure π. Then, the dynamical system $(\mathsf{X}^{\mathbb{N}},\mathscr{X}^{\otimes\mathbb{N}},\mathbb{P}_\pi,S)$ is ergodic.

PROOF. Let P be a Markov kernel that admits a unique invariant probability measure π and let $A\in\mathscr{I}=\{A\in\mathscr{X}^{\otimes\mathbb{N}}\,:\,\mathbf{1}_A=\mathbf{1}_A\circ S\}$.

Assume that $\mathbb{P}_\pi(A)>0$. We will show that $\mathbb{P}_\pi(A)=1$ and this will prove that the dynamical system $(\mathsf{X}^{\mathbb{N}},\mathscr{X}^{\otimes\mathbb{N}},\mathbb{P}_\pi,S)$ is ergodic.

Before diving into the proof, let us take a few minutes to analyse the situation... The quantity of interest is $\mathbb{P}_\pi(A)$ while the assumption is on π (it is the unique invariant probability measure for P). A first step is to relate $\mathbb{P}_\pi(A)$ to π... Using the decomposition with respect to the initial state, we can write

$$\mathbb{P}_\pi(A)=\int_{\mathsf{X}}\pi(\mathrm{d}x)\mathbb{E}_x[\mathbf{1}_A]=\pi(h_A)\,,\quad\text{where we have set } h_A(x)=\mathbb{E}_x[\mathbf{1}_A]. \tag{1.2}$$

The proof proceeds in two steps. We first establish some results on h_A and then deduce properties of $\mathbb{P}_\pi(A)$ by constructing, from π and A, another invariant probability measure π_A.

(i) **($h_A(X_n)$ does not depend on n, $\mathbb{P}_\pi-a.s.$).** To see this, first write for any $n\in\mathbb{N}$,

$$\begin{aligned}\mathbb{E}_\pi[|h_A(X_0)-\mathbf{1}_A|]&=\mathbb{E}_\pi[|h_A(X_0)-\mathbf{1}_A|\circ S^n]\\&=\mathbb{E}_\pi[|h_A(\underbrace{X_0\circ S^n}_{X_n})-\underbrace{\mathbf{1}_A\circ S^n}_{\mathbf{1}_A}|]=\mathbb{E}_\pi[|h_A(X_n)-\mathbf{1}_A|].\end{aligned} \tag{1.3}$$

We now show that the rhs tends to 0. First, we find another expression for $h_A(X_n)$. Define $\mathscr{F}_n=\sigma(X_0,\dots,X_n)$. Then, $\mathbb{P}_\pi-a.s.$,

$$h_A(X_n)=\mathbb{E}_{X_n}[\mathbf{1}_A]\overset{(1)}{=}\mathbb{E}_\pi\left[\mathbf{1}_A\circ S^n|\mathscr{F}_n\right]\overset{(2)}{=}\mathbb{E}_\pi\left[\mathbf{1}_A|\mathscr{F}_n\right],$$

where $\overset{(1)}{=}$ comes from the Markov property and $\overset{(2)}{=}$ from the fact that $A\in\mathscr{I}$. Fix some $k\in\mathbb{N}$ and let $B\in\mathscr{F}_k$. Then, for all $n\geqslant k$, we have $\mathbb{E}_\pi[\mathbf{1}_B|\mathscr{F}_n]=\mathbf{1}_B$ and thus,

$$\mathbb{E}_\pi[|h_A(X_n)-\mathbf{1}_B|]=\mathbb{E}_\pi[|\mathbb{E}_\pi[\mathbf{1}_A-\mathbf{1}_B|\mathscr{F}_n]|]\leqslant\mathbb{E}_\pi[\mathbb{E}_\pi[|\mathbf{1}_A-\mathbf{1}_B||\mathscr{F}_n]]=\mathbb{E}_\pi[|\mathbf{1}_B-\mathbf{1}_A|]$$

This implies, by using the triangle inequality and then taking the limsup,

$$\limsup_{n\to\infty}\mathbb{E}_\pi[|h_A(X_n)-\mathbf{1}_A|]\leqslant 2\mathbb{E}_\pi[|\mathbf{1}_B-\mathbf{1}_A|].$$

Now, according to the approximation lemma (Lemma 1.6), the rhs can be made arbitrarily small for a suitable choice of k and $B\in\mathscr{F}_k$. Therefore, $\lim_{n\to\infty}\mathbb{E}_\pi[|h_A(X_n)-\mathbf{1}_A|]=0$. Combining this limiting result with (1.3), we deduce that $\mathbb{E}_\pi[|h_A(X_n)-\mathbf{1}_A|]$ is a constant that tends to 0 as n tends to infinity. It is thus equal to 0 for all $n\in\mathbb{N}$, and we have

$$h_A(X_0)=h_A(X_n)=\mathbf{1}_A\,,\quad\mathbb{P}_\pi-a.s. \tag{1.4}$$

(ii) Now, define the probability measure π_A on $(\mathsf{X},\mathscr{X})$ by $\pi_A(f)=\frac{\mathbb{E}_\pi[h_A(X_0)f(X_0)]}{\mathbb{P}_\pi(A)}$ for any non-negative measurable function f on X. Then, using the Markov property, Equation (1.4) with $n=1$ and $X_1\overset{\mathscr{L}}{=}X_0$ under $\mathbb{P}_\pi$ (that is, they share the same distribution under $\mathbb{P}_\pi$), we get

$$\begin{aligned}\mathbb{E}_\pi\left[h_A(X_0)\times Pf(X_0)\right]&=\mathbb{E}_\pi\left[h_A(X_0)\times f(X_1)\right]=\mathbb{E}_\pi\left[h_A(X_1)\times f(X_1)\right]\\&=\mathbb{E}_\pi\left[h_A(X_0)\times f(X_0)\right],\end{aligned}$$

Therefore, $\pi_AP(f)=\pi_A(Pf)=\pi_A(f)$, showing that π_A is an invariant probability measure for P and thus, $\pi=\pi_A$. Then,

$$\mathbb{P}_\pi(A)=\int_{\mathsf{X}}\pi(\mathrm{d}x)\mathbb{P}_x(A)=\pi(h_A)=\pi_A(h_A)=\frac{\mathbb{E}_\pi\left[h_A(X_0)\times h_A(X_0)\right]}{\mathbb{P}_\pi(A)}.$$

Applying (1.4) again yields $\mathbb{P}_\pi(A)=\mathbb{E}_\pi\left[\mathbf{1}_A\times\mathbf{1}_A\right]/\mathbb{P}_\pi(A)=1$, which completes the proof of the theorem.

■

As a consequence, Birkhoff's theorem for dynamical systems, Theorem 1.4, yields

THEOREM 1.8 . Let P be a Markov kernel admitting a unique invariant probability measure π. Then, for all measurable functions h on $\mathsf{X}^{\mathbb{N}}$ such that $\mathbb{E}_\pi[|h|] < \infty$, we have

$$\lim_{n\to\infty} n^{-1} \sum_{k=0}^{n-1} h(X_{k:\infty}) = \mathbb{E}_\pi[h], \quad \mathbb{P}_\pi - a.s.$$

or equivalently,

$$\lim_{n\to\infty} n^{-1} \sum_{k=0}^{n-1} h \circ S^k = \mathbb{E}_\pi[h], \quad \mathbb{P}_\pi - a.s.$$

A particular case of Theorem 1.8 is when $h(X_{0:\infty}) = f(X_0)$. In this case, we have the following corollary.

COROLLARY 1.9 . Let P be a Markov kernel admitting a unique invariant probability measure π. Then, for all measurable functions f on X such that $\pi(|f|) = \int_{\mathsf{X}} \pi(\mathrm{d}x)|f(x)| < \infty$, we have

$$\lim_{n\to\infty} n^{-1} \sum_{k=0}^{n-1} f(X_k) = \pi(f), \quad \mathbb{P}_\pi - a.s. \tag{1.5}$$

The limiting result Equation (1.5) is nice but it holds $\mathbb{P}_\pi - a.s.$ We now try to overcome this issue. Under the assumptions of Corollary 1.9, set $A = \left\{\lim_{n\to\infty} n^{-1} \sum_{k=0}^{n-1} f(X_k) = \pi(f)\right\}$. Combining Theorem 1.8 with the decomposition in (1.2), $0 = \mathbb{P}_\pi(A^c) = \int_{\mathsf{X}} \pi(\mathrm{d}x)\mathbb{P}_x(A^c)$ and this implies that $\mathbb{P}_x(A^c) = 0$ (i.e. $\mathbb{P}_x(A) = 1$) for π-almost all $x \in \mathsf{X}$. Therefore, we finally get:

COROLLARY 1.10 . Let P be a Markov kernel admitting a unique invariant probability measure π. Then, for all measurable functions f on X such that $\pi(|f|) = \int_{\mathsf{X}} \pi(\mathrm{d}x)|f(x)| < \infty$, we have for π-almost all $x \in \mathsf{X}$,

$$\lim_{n\to\infty} n^{-1} \sum_{k=0}^{n-1} f(X_k) = \pi(f), \quad \mathbb{P}_x - a.s. \tag{1.6}$$

What about Metropolis-Hastings algorithms? Let $P^{\mathrm{MH}}_{\langle \pi, Q\rangle}$ denote the Metropolis-Hastings kernel with target π and proposal Q. When $Q(x,\mathrm{d}y) = q(x,y)\lambda(\mathrm{d}y)$ and $\pi(\mathrm{d}y) = \pi(y)\lambda(\mathrm{d}y)$, with $q > 0$ and $\pi > 0$, its acceptance probability and transition kernel are

$$\begin{aligned} \alpha(x,y) &= 1 \wedge \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}, \\ P^{\mathrm{MH}}_{\langle \pi, Q\rangle}(x,\mathrm{d}y) &= q(x,y)\alpha(x,y)\lambda(\mathrm{d}y) + \bar{\alpha}(x)\delta_x(\mathrm{d}y), \end{aligned} \tag{1.7}$$

where $\bar{\alpha}(x) = 1 - \int q(x,y)\alpha(x,y)\lambda(\mathrm{d}y)$. The detailed balance identity shows that π is invariant. Under these positivity assumptions, it is also the unique invariant probability measure, as shown in the proof below. Therefore, Equations (1.5) and (1.6) hold. In fact, we have an even stronger result.

THEOREM 1.11 . Let Q be a Markov kernel on $\mathsf{X}\times\mathscr{X}$ and $\pi\in M_1(\mathsf{X})$. Assume that $Q(x,dy)=q(x,y)\lambda(dy)$ and $\pi(dy)=\pi(y)\lambda(dy)$ where $q>0$, $\pi>0$ and λ is a σ-finite measure on $(\mathsf{X},\mathscr{X})$. Then, for the Markov chain $\{X_n : n\in\mathbb{N}\}$ with Markov kernel $P^{\mathrm{MH}}_{\langle\pi,Q\rangle}$, generated by the Metropolis-Hastings algorithm, the following holds for all initial distributions $\nu\in M_1(\mathsf{X})$ and all measurable functions f on X such that $\pi(|f|)=\int_{\mathsf{X}}\pi(dx)|f(x)|<\infty$,

$$\lim_{n\to\infty} n^{-1}\sum_{k=0}^{n-1} f(X_k)=\pi(f)\,,\quad \mathbb{P}_\nu-a.s \tag{1.8}$$

PROOF. Let $\nu\in M_1(\mathsf{X})$. Set $A=\left\{\lim_{n\to\infty} n^{-1}\sum_{k=0}^{n-1} f(X_k)=\pi(f)\right\}$. We first check uniqueness of the invariant probability measure. Since $q(x,y)\alpha(x,y)>0$, the measure $\mu P^{\mathrm{MH}}_{\langle\pi,Q\rangle}$ is equivalent to λ for every probability measure μ that is itself equivalent to λ, and it dominates λ for every probability measure μ. Thus every invariant probability measure dominates λ.

Suppose that π and another invariant probability measure differ. The positive and negative parts of their difference have the same nonzero mass and are invariant: applying $P^{\mathrm{MH}}_{\langle\pi,Q\rangle}$ to these parts gives a decomposition of the same signed measure with the same total mass, so it is its Jordan decomposition. After normalization, they are mutually singular invariant probability measures. This contradicts the fact that both dominate λ. Hence π is unique and, by Corollary 1.9, $0=\mathbb{P}_\pi(A^c)=\int_{\mathsf{X}}\pi(dx)h_{A^c}(x)=\pi(h_{A^c})$ where we have set $h_{A^c}(x)=\mathbb{E}_x[\mathbf{1}_{A^c}]$. Fix an arbitrary $x\in\mathsf{X}$. Since

$$\left(\lim_{n\to\infty} n^{-1}\sum_{k=0}^{n-1} f(X_k)=\pi(f)\right)\Leftrightarrow\left(\lim_{n\to\infty} n^{-1}\sum_{k=1}^{n} f(X_k)=\pi(f)\right),$$

the set $A\in\mathscr{I}$ and hence $A^c\in\mathscr{I}$ since $\mathscr{I}$ is a σ-field. Then, using the Markov property and $\mathbf{1}_{A^c}\circ S=\mathbf{1}_{A^c}$, we obtain

$$P^{\mathrm{MH}}_{\langle\pi,Q\rangle}h_{A^c}(x)=\mathbb{E}_x\left[\mathbb{E}_{X_1}[\mathbf{1}_{A^c}]\right]=\mathbb{E}_x\left[\mathbb{E}_x[\mathbf{1}_{A^c}\circ S|\mathscr{F}_1]\right]=\mathbb{E}_x[\mathbf{1}_{A^c}\circ S]=\mathbb{E}_x[\mathbf{1}_{A^c}]=h_{A^c}(x).$$

Combining this with (1.7) yields

$$h_{A^c}(x)=P^{\mathrm{MH}}_{\langle\pi,Q\rangle}h_{A^c}(x)=\int\frac{q(x,y)\alpha(x,y)}{\pi(y)}h_{A^c}(y)\pi(y)\lambda(dy)+\bar\alpha(x)h_{A^c}(x).$$

Since $q>0$ and $\pi>0$, we can easily check that $\alpha>0$ and $\bar\alpha(x)<1$ (check it carefully). The first term in the rhs is zero since $\pi(h_{A^c})=0$. Therefore, $(1-\bar\alpha(x))h_{A^c}(x)=0$ and since $\bar\alpha(x)<1$, we can conclude that $h_{A^c}(x)=\mathbb{E}_x[\mathbf{1}_{A^c}]=0$. Finally, x being arbitrary, we obtain

$$\mathbb{P}_\nu(A^c)=\int_{\mathsf{X}}\nu(dx)\mathbb{E}_x[\mathbf{1}_{A^c}]=0,$$

and this is equivalent to $\mathbb{P}_\nu\left(\lim n^{-1}\sum_{k=0}^{n-1} f(X_k)=\pi(f)\right)=1$. This completes the proof. ■

Theorem 1.11 is useful since the LLN holds $\mathbb{P}_\nu$-a.s. for all starting distributions ν, but it only concerns MH kernels. For a general kernel P, we have the following result.

THEOREM 1.12 . Let P be a Markov kernel on $\mathsf{X}\times\mathscr{X}$ that admits a unique invariant probability measure π. Assume in addition that for all bounded measurable functions h and all probability measures $\nu\in M_1(\mathsf{X})$,

$$\lim_{n\to\infty}\nu P^n h=\pi(h) \tag{1.9}$$

Then, for all initial distributions $\nu\in M_1(\mathsf{X})$ and all measurable functions f on X such that $\pi(|f|)=\int_{\mathsf{X}}\pi(dx)|f(x)|<\infty$,

$$\lim_{n\to\infty} n^{-1}\sum_{k=0}^{n-1} f(X_k)=\pi(f)\,,\quad \mathbb{P}_\nu-a.s \tag{1.10}$$

PROOF. We start as in the proof of Theorem 1.11. Let $\nu \in M_1(X)$. Set $A = \left\{\lim_{n\to\infty} n^{-1}\sum_{k=0}^{n-1} f(X_k) = \pi(f)\right\}$. Since π is the unique invariant probability measure for P, we have, by Corollary 1.9, $0 = \mathbb{P}_\pi(A^c) = \int_X \pi(dx) h_{A^c}(x) = \pi(h_{A^c})$ where we have set $h_{A^c}(x) = \mathbb{E}_x[\mathbf{1}_{A^c}]$. As in the proof of Theorem 1.11, $A^c \in \mathscr{I}$. Moreover,

$$\nu P^n h_{A^c} = \mathbb{E}_\nu[h_{A^c}(X_n)] = \mathbb{E}_\nu[\mathbb{E}_{X_n}[\mathbf{1}_{A^c}]] \overset{(1)}{=} \mathbb{E}_\nu[\mathbb{E}_\nu[\mathbf{1}_{A^c}\circ S^n|\mathscr{F}_n]] \overset{(2)}{=} \mathbb{E}_\nu[\mathbf{1}_{A^c}\circ S^n] \overset{(3)}{=} \mathbb{P}_\nu(A^c)$$

where $\overset{(1)}{=}$ follows from the Markov property, $\overset{(2)}{=}$ is the tower property, and $\overset{(3)}{=}$ follows from $\mathbf{1}_{A^c}\circ S^n = \mathbf{1}_{A^c}$, since $A^c \in \mathscr{I}$.

Combined with (1.9), we get

$$\mathbb{P}_\nu(A^c) = \lim_{n\to\infty} \nu P^n h_{A^c} = \pi(h_{A^c}) = \int \pi(dx)\mathbb{E}_x[\mathbf{1}_{A^c}] = \mathbb{P}_\pi(A^c) = 0$$

Finally $\mathbb{P}_\nu(A^c) = 0$, that is $\mathbb{P}_\nu(A) = 1$, which is equivalent to

$$\lim_{n\to\infty} n^{-1}\sum_{k=0}^{n-1} f(X_k) = \pi(f), \quad \mathbb{P}_\nu - a.s$$

■

Chapter 2

Geometric ergodicity

Time schedule (Note 1): 3H-Session

Let (X_n) be a Markov chain with Markov kernel P, and assume that P admits an invariant probability measure π. In this chapter, we are interested in identifying conditions under which one can obtain explicit bounds on the error between the marginal distribution of X_n and the target distribution π. To this end, we first introduce an appropriate notion of distance between probability measures, focusing in particular on the total variation distance and the V-norm. We then study the discrepancy between π and $P^n(x,\cdot)$, the distribution of the n-th iterate of the Markov kernel starting from an arbitrary point $x \in \mathsf{X}$, by deriving bounds on the total variation distance or V-norm between these two measures.

2.1 Total variation norm and coupling

We start with the notion of coupling between probability measures. If μ,ν are probability measures on $(\mathsf{X},\mathscr{X})$, then a coupling γ of (μ,ν) is a probability measure on the product space $(\mathsf{X}^2,\mathscr{X}^{\otimes 2})$ such that, if $(X,Y)\sim\gamma$, the marginals satisfy $X\sim\mu$ and $Y\sim\nu$.

Definition 2.1 . Let $(\mathsf{X},\mathscr{X})$ be a measurable space, and let $\mu,\nu\in \mathrm{M}_1(\mathsf{X})$ be probability measures. We define $\mathscr{C}(\mu,\nu)$, the set of couplings of (μ,ν), as follows:

$$\mathscr{C}(\mu,\nu)=\left\{\gamma\in \mathrm{M}_1(\mathsf{X}^2)\ :\ \forall A\in\mathscr{X},\gamma(A\times\mathsf{X})=\mu(A),\gamma(\mathsf{X}\times A)=\nu(A)\right\}$$

Any $\gamma\in\mathscr{C}(\mu,\nu)$ is called a coupling of (μ,ν).

To illustrate the definition of a coupling, consider first a simple example. One can construct a coupling of (μ,μ) by sampling $X\sim\mu$ and then setting $Y=X$. The joint distribution of (X,Y) is then a coupling of (μ,μ). This construction yields a maximally dependent coupling, since the two components are equal by design.

Alternatively, one may construct an independent coupling by drawing X and Y independently according to the same distribution μ. In this case, the joint distribution of (X,Y) is again a coupling of (μ,μ), but now with complete independence between the two coordinates.

These examples naturally raise the question of which coupling is the most relevant or useful. As we shall see in what follows, there is a substantial degree of freedom in the choice of a coupling, and the optimal choice depends on the problem under consideration. We will present several couplings that are particularly useful in practice, but there is no universal rule that applies in all situations.

Before going further into coupling techniques, we introduce the total variation norm and explain its fundamental connection with couplings.

Throughout this chapter, we assume that the diagonal $\Delta_{\mathsf{X}} = \{(x,x) : x \in \mathsf{X}\}$ belongs to $\mathscr{X} \otimes \mathscr{X}$, so that $\{X \neq Y\}$ is measurable for every coupling considered.

DEFINITION 2.2 . Let $(\mathsf{X}, \mathscr{X})$ be a measurable space, and let $\mu, \nu \in \mathrm{M}_1(\mathsf{X})$ be probability measures. We denote by $\mathsf{F}(\mathsf{X})$ the set of real-valued $\mathscr{X}$-measurable functions on X. The total variation norm between μ and ν, denoted by $\|\mu - \nu\|_{\mathrm{TV}}$, is defined by

$$\|\mu - \nu\|_{\mathrm{TV}} = 2 \sup \{|\mu(f) - \nu(f)| \ : \ f \in \mathsf{F}(\mathsf{X}), 0 \leqslant f \leqslant 1\} \tag{2.1}$$

$$= \int |\varphi_0 - \varphi_1|(x) \zeta(\mathrm{d}x) \tag{2.2}$$

$$= 2 \inf \{\mathbb{P}(X \neq Y) \ : \ (X,Y) \sim \gamma \text{ where } \gamma \in \mathscr{C}(\mu, \nu)\} \tag{2.3}$$

where $\mu(\mathrm{d}x) = \varphi_0(x)\zeta(\mathrm{d}x)$ and $\nu(\mathrm{d}x) = \varphi_1(x)\zeta(\mathrm{d}x)$.

Before proving that these different expressions of the total variation norm are indeed equivalent, let us make a few remarks.

(a) The reader might wonder why we can always write $\mu(\mathrm{d}x) = \varphi_0(x)\,\zeta(\mathrm{d}x)$ and $\nu(\mathrm{d}x) = \varphi_1(x)\,\zeta(\mathrm{d}x)$ for some suitably chosen measure ζ, and for measurable functions φ_0 and φ_1. Indeed, if $\mu, \nu \in \mathrm{M}_1(\mathsf{X})$, then by setting $\zeta = \mu + \nu$, we obtain the two implications

$$(\zeta(A) = 0) \Longrightarrow (\mu(A) = 0) \quad \text{and} \quad (\zeta(A) = 0) \Longrightarrow (\nu(A) = 0)\,.$$

Therefore, the measure ζ dominates both μ and ν. By the Radon–Nikodym theorem, the measures μ and ν admit densities with respect to ζ, which we denote by φ_0 and φ_1 in Definition 2.2.

(b) The first definition, (2.1), is expressed as a supremum over a class of test functions, while the last one is formulated as an infimum over coupling measures. These two characterizations can therefore be interpreted as a duality formula.

(c) The intermediate expression shows that the total variation norm can be written as an L_1-distance between the densities of the two distributions with respect to a common dominating measure.

(d) A direct consequence of the equivalence between these definitions of the total variation norm is the following. If f is a measurable function taking values in $[0,1]$, and if X, Y are random variables such that $(X,Y) \sim \gamma$ with $\gamma \in \mathscr{C}(\mu, \nu)$, then the coupling inequality

$$|\mu(f) - \nu(f)| \leqslant \mathbb{P}(X \neq Y)$$

holds. This inequality will be used repeatedly in what follows, and it highlights why coupling techniques are so powerful in practice.

PROOF. [of the equivalences in Definition 2.2] Let A, B, and C denote the quantities appearing in (2.1), (2.2), and (2.3), respectively. For any function f, $0 \leqslant f \leqslant 1$ if and only if the function $g = 2f - 1$ satisfies $|g| \leqslant 1$. This immediately implies

$$A = 2 \sup \{|\mu(f) - \nu(f)| \ : \ f \in \mathsf{F}(\mathsf{X}), 0 \leqslant f \leqslant 1\} = \sup \{|\mu(g) - \nu(g)| \ : \ g \in \mathsf{F}(\mathsf{X}), |g| \leqslant 1\}$$

Moreover, for any $g \in \mathsf{F}(\mathsf{X})$ such that $|g| \leqslant 1$,

$$|\mu(g) - \nu(g)| = \left| \int (\varphi_1 - \varphi_0)(x) g(x) \zeta(\mathrm{d}x) \right| \leqslant \int |\varphi_1 - \varphi_0|(x) \underbrace{|g(x)|}_{\leqslant 1} \zeta(\mathrm{d}x) = B$$

Therefore, $A \leqslant B$. Moreover, setting $g^*(x) = \operatorname{sign}(\varphi_0(x) - \varphi_1(x))$, we have $|g^*| = 1$, and hence

$$\begin{aligned} B &= \int |\varphi_0 - \varphi_1|(x)\zeta(\mathrm{d}x) = \int (\varphi_0(x) - \varphi_1(x))\, g^*(x)\zeta(\mathrm{d}x) \\ &= \mu(g^*) - \nu(g^*) \leqslant \sup\{|\mu(g) - \nu(g)| \,:\, g \in \mathsf{F}(\mathsf{X}), |g| \leqslant 1\} = A \end{aligned}$$

Thus, $\boxed{A = B}$. Now let $f \in (\mathsf{X})$ satisfy $0 \leqslant f \leqslant 1$, and let X, Y be random variables such that $(X,Y) \sim \gamma$ for some $\gamma \in \mathscr{C}(\mu,\nu)$. Then

$$|\mu(f) - \nu(f)| = |\mathbb{E}[f(X) - f(Y)]| = |\mathbb{E}\left[\{f(X) - f(Y)\}\mathbf{1}_{X \neq Y}\right]| \leqslant \mathbb{E}[\underbrace{|f(X) - f(Y)|}_{\leqslant 1}\mathbf{1}_{X \neq Y}] \leqslant \mathbb{P}(X \neq Y)$$

This shows that $\boxed{A \leqslant C}$. To finish the proof, we show that $C \leqslant B$ by exhibiting an **optimal coupling** of (μ,ν).

Define

$$\varepsilon = \int_{\mathsf{X}} \varphi_0 \wedge \varphi_1(x)\zeta(\mathrm{d}x) \quad \text{and} \quad \zeta'(\mathrm{d}x) = \frac{\varphi_0 \wedge \varphi_1(x)}{\varepsilon}\zeta(\mathrm{d}x)$$

It is straightforward to check that $\zeta' \in \mathrm{M}_1(\mathsf{X})$, and that $\mu(\mathrm{d}x) \geqslant \varepsilon\zeta'(\mathrm{d}x)$ and $\nu(\mathrm{d}x) \geqslant \varepsilon\zeta'(\mathrm{d}x)$. Therefore, there exist $\mu_1, \nu_1 \in \mathrm{M}_1(\mathsf{X})$ such that

$$\begin{aligned} \mu(\mathrm{d}x) &= \varepsilon\zeta'(\mathrm{d}x) + (1-\varepsilon)\mu_1(\mathrm{d}x) \\ \nu(\mathrm{d}x) &= \varepsilon\zeta'(\mathrm{d}x) + (1-\varepsilon)\nu_1(\mathrm{d}x) \end{aligned}$$

Define $\gamma(\mathrm{d}x\mathrm{d}y) = \varepsilon\zeta'(\mathrm{d}x)\delta_x(\mathrm{d}y) + (1-\varepsilon)\mu_1(\mathrm{d}x)\nu_1(\mathrm{d}y)$. It is clear that $\gamma \in \mathscr{C}(\mu,\nu)$. Let X, Y be random variables such that $(X,Y) \sim \gamma$. We can sample (X,Y) as follows: draw a Bernoulli random variable $U \sim \mathrm{Ber}(\varepsilon)$. If $U = 1$, draw $X \sim \zeta'$ and set $Y = X$. If $U = 0$, draw $X \sim \mu_1$ and $Y \sim \nu_1$ independently. Then

$$\begin{aligned} \mathbb{P}(X \neq Y) &= 1 - \mathbb{P}(X = Y) \leqslant 1 - \varepsilon = 1 - \int_{\mathsf{X}} \varphi_0 \wedge \varphi_1(x)\zeta(\mathrm{d}x) \\ &= \frac{1}{2}\int_{\mathsf{X}} \underbrace{\varphi_0 + \varphi_1(x) - 2\varphi_0 \wedge \varphi_1(x)}_{|\varphi_0 - \varphi_1(x)|}\, \zeta(\mathrm{d}x) \end{aligned}$$

This shows that $\boxed{C \leqslant B}$ and completes the proof. ■

EXERCISE 1 . Show that the supremum in (2.1) is attained for a suitable choice of f.

Another equivalent expression for the total variation distance is

$$\|\mu - \nu\|_{\mathrm{TV}} = 2\inf\{\gamma(\Delta) \,:\, \gamma \in \mathscr{C}(\mu,\nu)\} \tag{2.4}$$

where we have used the notation $\Delta(x,y) = \mathbf{1}_{x \neq y}$ (we also say that $\Delta(x,y)$ is the Hamming distance between x and y).

EXAMPLE 2.3 . See Figure 2.1. Let $\mu = \mathscr{N}(-1,1)$ and $\nu = \mathscr{N}(1,1)$. Let $X \sim \mathscr{N}(-1,1)$ and set $Y = X + 2$. Then, (X,Y) is a coupling of (μ,ν) but it is not the optimal coupling for the Hamming distance since $\mathbb{P}(X \neq Y) = 1$, whereas using Definition 2.2,

$$\|\mu - \nu\|_{\mathrm{TV}} = 2\left(1 - \int_{-\infty}^{\infty} \phi(x+1) \wedge \phi(x-1)\mathrm{d}x\right) = 2\left(1 - 2\int_1^{\infty} \frac{\mathrm{e}^{-u^2/2}}{\sqrt{2\pi}}\mathrm{d}u\right),$$

where ϕ is the density of the standard Gaussian distribution.

Before moving to geometric ergodicity, we define a generalisation of the total variation norm, namely the V-norm.

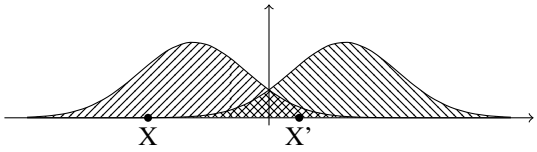


Figure 2.1: An example of coupling of two probability measures.

DEFINITION 2.4 . Let $(\mathsf{X}, \mathscr{X})$ be a measurable space, and let $\mu, \nu \in \mathrm{M}_1(\mathsf{X})$ be probability measures. Let $V : \mathsf{X} \to [1, \infty)$ be a measurable function, and assume that $\mu(V) + \nu(V) < \infty$. The V-norm between μ and ν, denoted by $\|\mu - \nu\|_{\mathrm{V}}$, is defined by

$$\begin{aligned}\|\mu - \nu\|_{\mathrm{V}} &= \sup \{|\mu(f) - \nu(f)| \ : \ f \in \mathsf{F}(\mathsf{X}), |f| \leqslant V\} \\ &= \int |\varphi_0 - \varphi_1|(x)\, V(x)\, \zeta(\mathrm{d}x) \\ &= \inf \left\{ \mathbb{E}\left[\{V(x) + V(y)\} 1_{\{X \neq Y\}} \right] \ : \ (X, Y) \sim \gamma \text{ where } \gamma \in \mathscr{C}(\mu, \nu) \right\}\end{aligned}$$

where $\mu(\mathrm{d}x) = \varphi_0(x)\zeta(\mathrm{d}x)$ and $\nu(\mathrm{d}x) = \varphi_1(x)\zeta(\mathrm{d}x)$.

Hence, the total variation norm is a particular case of the V-norm, corresponding to the choice $V \equiv 1$. We leave the proof of the equivalence between the various definitions of the V-norm to the reader, since it follows closely the arguments used to establish the analogous equivalences for the total variation norm.

From Definition 2.4, we obtain the following equivalent expression for the V-norm:

$$\|\mu - \nu\|_{\mathrm{V}} = 2 \inf \{\gamma(\bar{V}\Delta) \ : \ \gamma \in \mathscr{C}(\mu, \nu)\} \tag{2.5}$$

where we have set $\bar{V}(x, x') = [V(x) + V(x')]/2$. It is also useful to note that the total variation distance is dominated by the V-norm, that is, $\|\mu - \nu\|_{\mathrm{TV}} \leqslant \|\mu - \nu\|_{\mathrm{V}}$.

2.2 Geometric ergodicity

In what follows, we assume that there is **a measurable function** $V : \mathsf{X} \to [1, \infty)$ such that

(A1) [**Minorizing condition**] For every $d > 0$, there exist $\varepsilon_d > 0$ and a probability measure ν_d such that

$$\forall x \in C_d := \{V \leqslant d\}, \quad P(x, \cdot) \geqslant \varepsilon_d \nu_d(\cdot) \tag{2.6}$$

(A2) [**Drift condition**] There exist constants $(\lambda, b) \in (0, 1) \times \mathbb{R}^+$ such that, for all $x \in \mathsf{X}$,

$$PV(x) \leqslant \lambda V(x) + b$$

Typically, the function V is unbounded, although in specific situations it may also be bounded. Moreover, the level set $\{V \leqslant d\}$ is usually compact when the chain takes values in a topological space. Roughly

speaking, Assumption (A1) states that whenever x belongs to the set C_d, the transition measure $P(x,\cdot)$ is uniformly lower bounded by the non-trivial measure $\varepsilon_d \nu_d(\cdot)$.

In many applications, the state space is $\mathsf{X} = \mathbb{R}^n$, and the Markov kernel P is dominated by the Lebesgue measure, that is, $P(x,\mathrm{d}y) = p(x,y)\mathrm{d}y$. In this setting, we typically impose the minorization condition in the form $P(x,A) \geqslant \varepsilon_d \nu_d(A)$, where

$$\varepsilon_d = \int_{\mathsf{X}} \left[\inf_{x \in C_d} p(x,y) \right] \mathrm{d}y, \quad \nu_d(A) = \frac{\int_A \inf_{x \in C_d} p(x,y)\mathrm{d}y}{\varepsilon_d}.$$

In other words, it is sufficient to derive a uniform lower bound on the transition density $p(x,y)$ for $x \in C_d$. When C_d is compact, such a lower bound is often relatively easy to establish. In the terminology of Markov chains, whenever (2.6) holds, the set C_d is called a small set.

The drift condition in Assumption (A2) expresses that, on average, the drift function V is contracted by a factor λ, up to an additive constant b. Intuitively, this means that the Markov kernel P prevents the chain from spending too much time in regions where V takes very large values. Since moderate values of V typically correspond to bounded regions of the state space, this condition ensures that the chain does not escape to infinity too rapidly. It is therefore natural to expect such chains to exhibit good ergodic properties.

Before stating the main result, it is worth emphasizing that, in practice, there is no general recipe for choosing a drift function V satisfying Assumption (A2) for a given Markov kernel P. One usually has to experiment with several candidate functions. For instance, if $X_{k+1} = \alpha X_k + \varepsilon_k$, where (ε_k) are i.i.d. random variables and $\alpha \in (0,1)$, then assuming $\mathbb{E}[|\varepsilon_1|^r] < \infty$, a natural choice is $V(x) = |x|^r$. If instead $\mathbb{E}[\mathrm{e}^{\beta\varepsilon_1}] < \infty$, one may try $V(x) = \mathrm{e}^{\beta x}$. For Metropolis–Hastings algorithms, it is also common to consider negative powers of the target density. In all cases, however, the choice of V is highly model dependent, and, in some sense, this step provides an opportunity for a certain amount of creativity.

We now show that Assumptions (A1) and (A2) imply that the Markov kernel P is geometrically ergodic in the following sense.

THEOREM 2.5 . [Geometric ergodicity] Assume (A1) and (A2) for some measurable function $V \geqslant 1$. Then there exist constants $\rho \in (0,1)$ and $\alpha > 0$ such that, for all $x, x' \in \mathsf{X}$ and all $n \in \mathbb{N}$,

$$\left\| P^n(x,\cdot) - P^n(x',\cdot) \right\|_V \leqslant \alpha \rho^n \left[V(x) + V(x') \right].$$

REMARK 2.6 . Assume that there exists a constant $\varepsilon > 0$ and a probability measure ν such that, for all $x \in \mathsf{X}$, $P(x,\cdot) \geqslant \varepsilon \nu(\cdot)$. In that case, (A1) and (A2) hold with the constant function $V(x) = 1$, and Theorem 2.5 shows that

$$\left\| P^n(x,\cdot) - P^n(x',\cdot) \right\|_{\mathrm{TV}} \leqslant 2\alpha\rho^n.$$

for some constants $(\rho, \alpha) \in (0,1) \times \mathbb{R}^+$. Such a Markov chain is usually called uniformly ergodic.

EXERCISE 2 . Assume that there exists a constant $\varepsilon > 0$ and a probability measure ν such that, for all $x \in \mathsf{X}$, $P(x,\cdot) \geqslant \varepsilon \nu(\cdot)$. Show that there exists a unique invariant probability measure π and that

$$\| P^n(x,\cdot) - \pi \|_{\mathrm{TV}} \leqslant 2(1-\varepsilon)^n$$

The proof proceeds in several steps. To bound $\| P^n(x,\cdot) - P^n(x',\cdot) \|_V$, we construct a bivariate Markov chain (X_k, X'_k) such that the first component (X_k) evolves marginally according to the Markov kernel P started at x, while the second component (X'_k) evolves marginally according to P started at x'. This coupling allows us to study the two marginal distributions simultaneously and control their discrepancy through their joint evolution. We now describe the construction in more detail.

In what follows, we choose d sufficiently large so that

$$\boxed{\bar{\lambda} := \lambda + \frac{2b}{1+d} < 1} \tag{2.7}$$

Definition of the joint kernel $\bar{P}$

Define $Q(x_k, \mathrm{d}x_{k+1}) = \frac{P(x_k,\mathrm{d}x_{k+1}) - \varepsilon_d \nu_d(\mathrm{d}x_{k+1})}{1-\varepsilon_d}$ and set

$$\begin{aligned}\bar{P}((x_k,x'_k),\mathrm{d}x_{k+1}\mathrm{d}x'_{k+1}) &= \mathbf{1}_{x_k=x'_k} P(x_k,\mathrm{d}x_{k+1})\delta_{x_{k+1}}(x'_{k+1}) \\ &+ \mathbf{1}_{x_k\neq x'_k}\mathbf{1}_{(x_k,x'_k)\notin C_d^2}\left[P(x_k,\mathrm{d}x_{k+1})P(x'_k,\mathrm{d}x'_{k+1})\right] \\ &+ \mathbf{1}_{x_k\neq x'_k}\mathbf{1}_{(x_k,x'_k)\in C_d^2}\left[\varepsilon_d\nu_d(\mathrm{d}x_{k+1})\delta_{x_{k+1}}(x'_{k+1}) + (1-\varepsilon_d)Q(x_k,\mathrm{d}x_{k+1})Q(x'_k,\mathrm{d}x'_{k+1})\right]\end{aligned}$$

In fact, $\bar{P}$ is a Markov kernel on $\mathsf{X}^2 \times \mathscr{X}^{\otimes 2}$, and it is straightforward to check that

$$\bar{P}((x,x'),\cdot) \in \mathscr{C}(P(x,\cdot),P(x',\cdot)) \tag{2.8}$$

By induction, this implies that, for any $n \in \mathbb{N}$,

$$\bar{P}^n((x,x'),\cdot) \in \mathscr{C}(P^n(x,\cdot),P^n(x',\cdot)) \tag{2.9}$$

Interpretation of the joint kernel $\bar{P}$

Set $\bar{X}_k = (X_k, X'_k)$ and $\bar{C}_d = C_d \times C_d$. If $(\bar{X}_k)_{k\in\mathbb{N}}$ is a Markov chain with kernel $\bar{P}$, its transition from $\bar{X}_k = (x_k,x'_k)$ to $\bar{X}_{k+1} = (X_{k+1}, X'_{k+1})$ is defined as follows:

- If $x_k = x'_k$, draw $X_{k+1} \sim P(x_k,\cdot)$ and set $X'_{k+1} = X_{k+1}$.
- Otherwise,
  - If $(x_k,x'_k) \notin \bar{C}_d$, then
    - Draw independently $X_{k+1} \sim P(x_k,\cdot)$ and $X'_{k+1} \sim P(x'_k,\cdot)$
  - If $(x_k,x'_k) \in \bar{C}_d$, then
    - Draw $U \sim \mathrm{Ber}(\varepsilon_d)$.
    - If $U = 1$, draw $X_{k+1} \sim \nu_d$ and set $X'_{k+1} = X_{k+1}$.
    - If $U = 0$, draw independently $X_{k+1} \sim Q(x_k,\cdot)$ and $X'_{k+1} \sim Q(x'_k,\cdot)$.
- Set $\bar{X}_{k+1} = (X_{k+1}, X'_{k+1})$.

Thus, the bivariate Markov chain $(\bar{X}_k)_{k\in\mathbb{N}} = (X_k, X'_k)_{k\in\mathbb{N}}$ attempts to couple its two components with probability ε_d each time it enters $\bar{C}_d$. Once they are coupled (that is, when $X_k = X'_k$), they remain together forever: $X_n = X'_n$ for all $n \geqslant k$.

Some useful properties of $\bar{P}$

The following inequalities are immediate:

1. Let $\Delta(x,x') = \mathbf{1}_{x\neq x'}$. Then
   (a) if $(x,x') \in \bar{C}_d$, $\bar{P}\Delta(x,x') \leqslant (1-\varepsilon_d)\Delta(x,x')$
   (b) if $(x,x') \notin \bar{C}_d$, $\bar{P}\Delta(x,x') \leqslant \Delta(x,x')$
2. Setting $\bar{V}(x,x') = (V(x)+V(x'))/2$, we have $\bar{P}\bar{V}(x,x') = 2^{-1}(PV(x)+PV(x')) \leqslant \lambda\bar{V}(x,x')+b$. This implies
   (a) if $(x,x') \in \bar{C}_d$, $\bar{P}\bar{V}(x,x') \leqslant (\lambda+b)\bar{V}(x,x')$
   (b) if $(x,x') \notin \bar{C}_d$, $\bar{P}\bar{V}(x,x') \leqslant \underbrace{\left(\lambda + \frac{2b}{1+d}\right)}_{\tilde{\lambda}}\bar{V}(x,x')$

We now have all the tools needed to prove Theorem 2.5.

PROOF. [of Theorem 2.5] For any $\eta > 0$, define

$$\rho_\eta = \max\left(\frac{(\lambda+b)d+\eta(1-\varepsilon)}{d+\eta}, \frac{\bar{\lambda}+\eta}{1+\eta}\right) \tag{2.10}$$

Since $\lim_{\eta\to\infty}\frac{(\lambda+b)d+\eta(1-\varepsilon)}{d+\eta} = 1-\varepsilon$, we may choose $\eta^\star$ such that $\rho_{\eta^\star} < 1$. This ensures that the contraction factor $\rho_{\eta^\star}$ is strictly less than 1, which is crucial for proving geometric convergence.

Set

$$W = (\bar{V}+\eta^\star)\Delta.$$

We first show that, for any $x, x' \in \mathsf{X}$,

$$\bar{P}W(x,x') \leqslant \rho_{\eta^\star} W(x,x') \tag{2.11}$$

By definition of the Markov kernel $\bar{P}$, we have $\bar{P}W(x,x) = 0 = \rho_{\eta^\star}W(x,x)$, so it suffices to prove (2.11) for $\boxed{x \neq x'}$.

Now, for any $x, x' \in \mathsf{X}$ such that $x \neq x'$, using the fact that $\Delta \leqslant 1$, we get

$$\begin{aligned}\bar{P}W(x,x') &\leqslant [\bar{P}\bar{V} + \eta^\star \bar{P}\Delta](x,x') \\ &\leqslant \begin{cases} [(\lambda+b)\bar{V} + \eta^\star(1-\varepsilon)](x,x') & \text{if } (x,x') \in \bar{C}_d \\ \left[\bar{\lambda}\bar{V} + \eta^\star\right](x,x') & \text{if } (x,x') \notin \bar{C}_d \end{cases}\end{aligned}$$

The right-hand side can be bounded by considering two cases:

Case 1: $(x,x') \in \bar{C}_d$

Observe that the function $v \mapsto \frac{(\lambda+b)v+\eta^\star(1-\varepsilon)}{v+\eta^\star}$ is monotone. It equals $1-\varepsilon < 1$ at $v = 0$ and converges to $\lambda + b > 1$ as $v \to \infty$, implying that the function is non-decreasing. Hence, for $(x,x') \in \bar{C}_d$ with $x \neq x'$,

$$\frac{(\lambda+b)\bar{V}(x,x')+\eta^\star(1-\varepsilon)}{\bar{V}(x,x')+\eta^\star} \leqslant \frac{(\lambda+b)d+\eta^\star(1-\varepsilon)}{d+\eta^\star} \leqslant \rho_{\eta^\star}$$

where the last inequality follows directly from (2.10).

Case 2: $(x,x') \notin \bar{C}_d$

Similarly, the function $v \mapsto \frac{\bar{\lambda}v+\eta^\star}{v+\eta^\star}$ is monotone. It equals 1 at $v = 0$ and converges to $\bar{\lambda} < 1$ as $v \to \infty$, so it is non-increasing. Since $\bar{V} \geqslant 1$, we have for $(x,x') \notin \bar{C}_d$ with $x \neq x'$

$$\frac{\bar{\lambda}\bar{V}(x,x')+\eta^\star}{\bar{V}(x,x')+\eta^\star} \leqslant \frac{\bar{\lambda}+\eta^\star}{1+\eta^\star} \leqslant \rho_{\eta^\star}$$

again using (2.10).

Combining the two cases, we have for any $x \neq x'$,

$$\bar{P}W(x,x') \leqslant \rho_{\eta^\star}\left[\bar{V}(x,x')+\eta^\star\right] = \rho_{\eta^\star}W(x,x')$$

where we used $\Delta(x,x') = 1$ in the last equality. Therefore, (2.11) holds for all $\boxed{x,x' \in \mathsf{X}}$.

Now, for any $x, x' \in \mathsf{X}$,

$$\left\|P^n(x,\cdot) - P^n(x',\cdot)\right\|_{\mathrm{V}} \overset{(1)}{\leqslant} 2\bar{P}^n[\bar{V}\Delta](x,x') \leqslant 2\bar{P}^n W(x,x') \overset{(2)}{\leqslant} 2\rho_{\eta^\star}^n W(x,x') \overset{(3)}{\leqslant} \alpha\rho_{\eta^\star}^n(V(x)+V(x'))$$

Here, $\overset{(1)}{\leqslant}$ follows from (2.9) and (2.5), $\overset{(2)}{\leqslant}$ from (2.11), and $\overset{(3)}{\leqslant}$ holds by setting $\alpha = (1+\eta^\star)/2$. This chain of inequalities establishes the desired geometric convergence in the V-norm.

■

COROLLARY 2.7 . Assume that (A1) and (A2) hold for some measurable function $V \geqslant 1$. Then, the Markov kernel P admits a unique invariant probability measure π. Moreover, $\pi(V) < \infty$ and there exist constants $(\rho,\alpha) \in (0,1)\times\mathbb{R}^+$ such that for all $\mu \in \mathrm{M}_1(\mathsf{X})$ with $\mu(V) < \infty$ and all $n \in \mathbb{N}$,

$$\|\mu P^n - \pi\|_{\mathrm{V}} \leqslant \alpha\rho^n\mu(V).$$

PROOF. For any $\mu,\nu \in \mathrm{M}_1(\mathsf{X})$ and any $h \in \mathsf{F}(\mathsf{X})$ such that $|h| \leqslant V$, we have, using Theorem 2.5,

$$|\mu P^n h - \nu P^n h| = |\int_{\mathsf{X}^2} \mu(\mathrm{d}x)\nu(\mathrm{d}y)[P^n h(x) - P^n h(y)]| \leqslant \int_{\mathsf{X}^2} \mu(\mathrm{d}x)\nu(\mathrm{d}y)|[P^n h(x) - P^n h(y)]| \leqslant \alpha\rho^n[\mu(V)+\nu(V)]$$

Thus,

$$\|\mu P^n - \nu P^n\|_{\mathrm{V}} \leqslant \alpha\rho^n[\mu(V)+\nu(V)] \tag{2.12}$$

Taking $\mu = \delta_x$ and $\nu = P(x,\cdot)$ gives, for every $x \in \mathsf{X}$,

$$\left\|P^n(x,\cdot) - P^{n+1}(x,\cdot)\right\|_{\mathrm{TV}} \leqslant \left\|P^n(x,\cdot) - P^{n+1}(x,\cdot)\right\|_{\mathrm{V}} \leqslant \alpha\rho^n[V(x)+PV(x)] \leqslant \alpha\rho^n[(1+\lambda)V(x)+b]$$

This implies that $\{P^n(x,\cdot)\}$ is a Cauchy sequence. Since $(\mathrm{M}_1(\mathsf{X}), \|\cdot\|_{\mathrm{TV}})$ is complete, it converges to a limit $\pi \in \mathrm{M}_1(\mathsf{X})$. For every $x \in \mathsf{X}$ and every $h \in \mathsf{F}(\mathsf{X})$ such that $|h| \leqslant 1$, we also have $|Ph| \leqslant 1$, and therefore

$$\pi(Ph) = \lim_{n\to\infty} P^n(Ph)(x) = \lim_{n\to\infty} P^{n+1}h(x) = \pi(h)$$

showing that π is P-**invariant**. We now prove **uniqueness** of the invariant probability measure. To see this, note that π does not depend on the choice of x. Indeed, taking $\mu = \delta_x$ and $\nu = \delta_{x'}$ in (2.12) gives $\lim_{n\to\infty} \|P^n(x,\cdot) - P^n(x',\cdot)\|_{\mathrm{TV}} = 0$. Therefore, for every $x \in \mathsf{X}$, $\lim_{n\to\infty} P^n h(x) = \pi(h)$. Let π' be an invariant probability measure for P. Then

$$\pi'(h) = \pi' P^n(h) = \int \pi'(\mathrm{d}x) \underbrace{P^n h(x)}_{\to \pi(h)} \to_{n\to\infty} \pi(h)$$

where the convergence follows from the dominated convergence theorem. Since $PV \leqslant \lambda V + b$, induction gives, for all $n \in \mathbb{N}$,

$$P^n V(x) \leqslant \lambda^n V(x) + b\left(\sum_{k=0}^{n-1} \lambda^k\right) \leqslant \lambda^n V(x) + \frac{b}{1-\lambda}$$

Therefore, for any $M > 0$, Jensen's inequality applied to the convex function $u \mapsto u \wedge M$ gives $P^n(V \wedge M)(x) \leqslant (P^n V(x)) \wedge M \leqslant \left(\lambda^n V(x) + \frac{b}{1-\lambda}\right) \wedge M$. We then integrate with respect to π and use $\pi = \pi P^n$:

$$\pi(V \wedge M) = \pi P^n(V \wedge M) \leqslant \int \pi(\mathrm{d}x)\left(\lambda^n V(x) + \frac{b}{1-\lambda}\right) \wedge M$$

Letting n tend to infinity and applying the dominated convergence theorem gives $\pi(V \wedge M) \leqslant \frac{b}{1-\lambda} \wedge M$. Then, letting M tend to infinity, we obtain $\pi(V) \leqslant b/(1-\lambda) < \infty$. To complete the proof, apply (2.12) with $\nu = \pi$ to obtain

$$\|\mu P^n - \underbrace{\pi P^n}_{\pi}\|_{\mathrm{TV}} \leqslant \alpha\rho^n[\mu(V)+\pi(V)] \leqslant \alpha'\rho^n\mu(V)$$

where we have set $\alpha' = \alpha[1+\pi(V)] < \infty$. ■

Chapter 3

Central limit theorems

Time schedule (Note 2): 3H-Session

3.1 The Poisson Equation

3.1.1 Definition

We start with a general definition of a Poisson equation.

DEFINITION 3.1 . For a given measurable function h such that $\pi|h| < \infty$, the Poisson equation is defined by

$$\hat{h} - P\hat{h} = h - \pi(h) \tag{3.1}$$

A solution to the Poisson equation (3.1) is a function $\hat{h}$ such that $P|\hat{h}|(x) < \infty$ for all $x \in \mathsf{X}$ and for all $x \in \mathsf{X}$, $\hat{h}(x) - P\hat{h}(x) = h(x) - \pi(h)$.

The following result holds under assumptions (A1) and (A2).

THEOREM 3.2 . Assume (A1) and (A2) hold for some measurable function $V \geqslant 1$. Then, for any measurable function h such that $|h| \leqslant V$, the function

$$\hat{h} = \sum_{n=0}^{\infty} \{P^n h - \pi(h)\} \tag{3.2}$$

is well-defined. Moreover, $\hat{h}$ is a solution of the Poisson equation associated with h and there exists a constant γ such that, for all $x \in \mathsf{X}$,

$$|\hat{h}(x)| \leqslant \gamma V(x)$$

PROOF. To see the existence of a solution to the Poisson equation under (A1) and (A2), note that by Theorem 2.5, $\sum_{n=0}^{\infty} \{P^n h(x) - \pi(h)\}$ converges for any $|h| \leqslant V$ and we can thus define

$$\hat{h}(x) = \sum_{n=0}^{\infty} \{P^n h(x) - \pi(h)\}$$

Then,

$$P\hat{h}(x) = \sum_{n=1}^{\infty} \{P^n h(x) - \pi(h)\}$$

which immediately shows (3.1). Moreover, setting $\hat{h}$ as in (3.2), Theorem 2.5 shows that for all $x \in \mathsf{X}$,

$$|\hat{h}(x)| \leqslant \frac{\alpha}{1-\rho} V(x)$$

■

3.1.2 Poisson equation and martingales

The Poisson equation is useful because it allows us to link quantities of interest for our Markov chain to a suitably chosen martingale. We can then apply limit theorems for martingales and transfer their implications to our Markov chain.

We begin with a review of martingales.

A review of martingales

Let $(M_n)_{n\in\mathbb{N}}$ be a sequence of random variables on the same probability space $(\Omega, \mathscr{F}, \mathbb{P})$, and let $(\mathscr{F}_n)_{n\in\mathbb{N}}$ be a filtration (i.e., $\mathscr{F}_n \subset \mathscr{F}_{n+1} \subset \mathscr{F}$ for all $n \in \mathbb{N}$). We say that $(M_n)_{n\in\mathbb{N}}$ is a $(\mathscr{F}_n)$-martingale if each M_n is integrable and, for all $n \geqslant 1$,

$$\mathbb{E}[M_n | \mathscr{F}_{n-1}] = M_{n-1}$$

The *increment process* of the martingale is by definition $(M_{n+1} - M_n)_{n\in\mathbb{N}}$.

The following CLT result holds for martingales whose increment process is stationary and ergodic. It is stated without proof.

THEOREM 3.3 . If a sequence $(M_n)_{n\in\mathbb{N}}$ is a $(\mathscr{F}_n)$-martingale whose increment process is stationary and ergodic, with square-integrable increments, then

$$n^{-1/2} M_n \overset{\mathscr{L}^{\mathbb{P}}}{\rightsquigarrow} \mathscr{N}\left(0, \mathbb{E}[(M_1 - M_0)^2]\right)$$

Link with martingales

Define

$$S_n(h) = \sum_{k=0}^{n-1} \{h(X_k) - \pi(h)\}$$

The solution to the Poisson equation allows us to express $S_n(h)$ in terms of a martingale:

$$S_n(h) = M_n(\hat{h}) + \hat{h}(X_0) - \hat{h}(X_n) \tag{3.3}$$

where

$$M_n(\hat{h}) = \sum_{k=1}^{n} \left\{\hat{h}(X_k) - P\hat{h}(X_{k-1})\right\} \tag{3.4}$$

Note that $\left\{M_n(\hat{h})\right\}_{n\in\mathbb{N}}$ is a $(\mathscr{F}_k)$-martingale, where $\mathscr{F}_k = \sigma(X_0, \ldots, X_k)$, since

$$\mathbb{E}[M_n(\hat{h}) | \mathscr{F}_{n-1}] - M_{n-1}(h) = \mathbb{E}[\hat{h}(X_n) - P\hat{h}(X_{n-1}) | \mathscr{F}_{n-1}] = P\hat{h}(X_{n-1}) - P\hat{h}(X_{n-1}) = 0$$

This connection with martingales allows us to derive laws of large numbers (LLNs) and central limit theorems (CLTs) for our Markov chain from limit theorems for martingales. Since we have already studied LLNs using several approaches in previous chapters, we focus here only on CLTs.

3.1.3 Central Limit theorems

THEOREM 3.4 . Let P be a Markov kernel with a unique invariant probability measure π. Let $h \in \mathsf{L}_2(\pi)$. Assume that there exists a solution $\boxed{\hat{h} \in \mathsf{L}_2(\pi)}$ to the Poisson equation $\hat{h} - P\hat{h} = h - \pi(h)$. Then

$$n^{-1/2} \sum_{k=0}^{n-1} \{h(X_k) - \pi(h)\} \overset{\mathscr{L}_{\mathbb{P}_\pi}}{\rightsquigarrow} \mathcal{N}(0, \sigma_\pi^2(h)) \,,$$

where

$$\sigma_\pi^2(h) = \mathbb{E}_\pi[\{\hat{h}(X_1) - P\hat{h}(X_0)\}^2] \tag{3.5}$$

PROOF. Without loss of generality, we assume $\pi(h) = 0$. The sequence $(M_n(\hat{h}))_{n\in\mathbb{N}}$ defined in (3.4) is such that

$$\begin{aligned}\mathbb{E}_\pi[(M_n(\hat{h}) - M_{n-1}(\hat{h}))^2] = \mathbb{E}_\pi[\{\hat{h}(X_1) - P\hat{h}(X_0)\}^2] &\leqslant 2\mathbb{E}_\pi[\hat{h}^2(X_1) + (P\hat{h}(X_0))^2] \\ &= 2\left[\pi(\hat{h}^2) + \pi((P\hat{h})^2)\right] \overset{(1)}{\leqslant} 2\left[\pi(\hat{h}^2) + \underbrace{\pi P}_{\pi}(\hat{h})^2\right] = 2\pi(\hat{h}^2) < \infty\end{aligned}$$

where $\overset{(1)}{\leqslant}$ follows from the Cauchy-Schwarz inequality. Therefore, the sequence $(M_n(\hat{h}))_{n\in\mathbb{N}}$ is a martingale with stationary, square-integrable increments under $\mathbb{P}_\pi$. By Theorem 3.3, we have

$$n^{-1/2} M_n(\hat{h}) \overset{\mathscr{L}_{\mathbb{P}_\pi}}{\rightsquigarrow} \mathcal{N}\left(0, \mathbb{E}_\pi[\{\hat{h}(X_1) - P\hat{h}(X_0)\}^2]\right) \,. \tag{3.6}$$

Since the Markov chain $(X_k)_{k\in\mathbb{N}}$ is stationary under $\mathbb{P}_\pi$, we have $\mathbb{E}_\pi[|\hat{h}(X_0) + \hat{h}(X_n)|] \leqslant 2\pi(|\hat{h}|)$ which implies that

$$n^{-1/2}\{\hat{h}(X_0) + \hat{h}(X_n)\} \overset{\mathbb{P}_\pi - prob}{\longrightarrow} 0 \,.$$

Combining this with (3.6) and (3.3), and applying Slutsky's lemma, we obtain

$$n^{-1/2} \sum_{k=0}^{n-1} h(X_k) \overset{\mathscr{L}_{\mathbb{P}_\pi}}{\rightsquigarrow} \mathcal{N}(0, \sigma_\pi^2(h))$$

■

THEOREM 3.5 . Assume that (A1) and (A2) hold for some function V. Then, for every measurable function h such that $|h|^2 \leqslant V$,

$$n^{-1/2} \sum_{k=0}^{n-1} \{h(X_k) - \pi(h)\} \overset{\mathscr{L}_{\mathbb{P}_\pi}}{\rightsquigarrow} \mathcal{N}(0, \sigma_\pi^2(h)) \,,$$

where

$$\sigma_\pi^2(h) = \mathbb{E}_\pi[\{\hat{h}(X_1) - P\hat{h}(X_0)\}^2] \tag{3.7}$$

and $\hat{h}$ is defined as in (3.2).

PROOF. Assume that (A1) and (A2) hold for some function V. Then, (A1) also holds with V replaced by $V^{1/2}$. Moreover, since $PV \leqslant \lambda V + b$, the Cauchy-Schwarz inequality gives

$$P(V^{1/2}) \leqslant (PV)^{1/2} \leqslant (\lambda V + b)^{1/2} \leqslant \lambda^{1/2} V^{1/2} + b^{1/2}$$

Finally, (A1) and (A2) hold for the function $V^{1/2}$. We can therefore apply Theorem 3.2 with V replaced by $V^{1/2}$. Then, for all $h \leqslant V^{1/2}$, the function $\hat{h}$ defined by (3.2) is a solution to the Poisson equation, and there exists a

constant $\gamma > 0$ such that $\hat{h} \leqslant \gamma V^{1/2}$. This implies that $\pi(\hat{h}^2) \leqslant \gamma\pi(V) < \infty$ by Corollary 2.7. Therefore, $\hat{h} \in \mathsf{L}_2(\pi)$, and Theorem 3.4 applies. This completes the proof. ■

Under the assumptions of Theorem 3.5, the CLT holds under $\mathbb{P}_\pi$. Moreover, this result extends to $\mathbb{P}_\nu$ for any probability measure ν in $\mathsf{M}_1(\mathsf{X})$.

This extension follows directly from the following proposition.

PROPOSITION 3.6 . Let P be a Markov kernel on $(\mathsf{X}, \mathscr{X})$ with an invariant probability measure π. Assume that there exists a constant $\sigma^2 \in (0,\infty)$ and a distribution ξ such that

$$n^{-1/2}\sum_{k=0}^{n-1}\{h(X_k) - \pi(h)\} \overset{\mathscr{L}_{\mathbb{P}_\pi}}{\rightsquigarrow} \mathscr{N}(0,\sigma^2)\,,$$
$$\lim_{n\to\infty} \|\xi P^n - \pi\|_{\mathrm{TV}} = 0$$

Then

$$n^{-1/2}\sum_{k=0}^{n-1}\{h(X_k) - \pi(h)\} \overset{\mathscr{L}_{\mathbb{P}_\xi}}{\rightsquigarrow} \mathscr{N}(0,\sigma^2)\,.$$

PROOF.

Without loss of generality, assume that $\pi(h) = 0$. We aim to show that

$$\limsup_{n\to\infty} \left|\mathbb{E}_\xi\left[e^{iu\left\{n^{-1/2}\sum_{k=0}^{n-1} h(X_k)\right\}}\right] - e^{-\sigma^2/2}\right| = 0. \tag{3.8}$$

Fix $p > 0$. For any $n \geqslant p$, decompose the error as

$$\left|\mathbb{E}_\xi\left[e^{iu\left\{n^{-1/2}\sum_{k=0}^{n-1} h(X_k)\right\}}\right] - e^{-\sigma^2/2}\right| \leqslant A_n + B_n + C_n + D_n,$$

where

$$A_n = \mathbb{E}_\xi\left[\left|e^{iu\left\{n^{-1/2}\sum_{k=0}^{n-1} h(X_k)\right\}} - e^{iu\left\{n^{-1/2}\sum_{k=p}^{n-1} h(X_k)\right\}}\right|\right],$$
$$B_n = \left|\mathbb{E}_\xi\left[e^{iu\left\{n^{-1/2}\sum_{k=p}^{n-1} h(X_k)\right\}}\right] - \mathbb{E}_\pi\left[e^{iu\left\{n^{-1/2}\sum_{k=p}^{n-1} h(X_k)\right\}}\right]\right|,$$
$$C_n = \mathbb{E}_\pi\left[\left|e^{iu\left\{n^{-1/2}\sum_{k=p}^{n-1} h(X_k)\right\}} - e^{iu\left\{n^{-1/2}\sum_{k=0}^{n-1} h(X_k)\right\}}\right|\right],$$
$$D_n = \left|\mathbb{E}_\pi\left[e^{iu\left\{n^{-1/2}\sum_{k=0}^{n-1} h(X_k)\right\}}\right] - e^{-\sigma^2/2}\right|.$$

By assumption, $\lim_{n\to\infty} D_n = 0$. Moreover,

$$A_n = \mathbb{E}_\xi\left[\left|e^{iu\left\{n^{-1/2}\sum_{k=0}^{p-1} h(X_k)\right\}} - 1\right|\right] \xrightarrow[n\to\infty]{} 0,$$

by the dominated convergence theorem. The same argument gives $\lim_{n\to\infty} C_n = 0$.

Next, define

$$\phi(x) = \mathbb{E}_x\left[e^{iu\left\{n^{-1/2}\sum_{\ell=0}^{n-p-1} h(X_\ell)\right\}}\right].$$

With this notation, the Markov property allows us to write B_n as

$$B_n = |\mathbb{E}_\xi[\phi(X_p)] - \mathbb{E}_\pi[\phi(X_p)]|.$$

Writing $\phi = \mathscr{R}e\,(\phi) + i\mathscr{I}m\,(\phi)$ and noting that $|\phi| \leqslant 1$, we obtain

$$B_n \leqslant 2\,\|\xi P^p - \pi\|_{\mathrm{TV}}\,.$$

Finally, we obtain

$$\limsup_{n\to\infty} \left|\mathbb{E}_\xi\left[e^{iu\left\{n^{-1/2}\sum_{k=0}^{n-1} h(X_k)\right\}}\right] - e^{-\sigma^2/2}\right| \leqslant 2\,\|\xi P^p - \pi\|_{\mathrm{TV}}\,.$$

Since p is arbitrary and $\lim_{p\to\infty} \|\xi P^p - \pi\|_{\mathrm{TV}} = 0$, this proves (3.8) and completes the proof. ■

Chapter 4

Spectral theory

Time schedule (Note 3): Last session.

4.1 Markov operator

We define $L_2(\pi)$ as the set of measurable functions f on X such that $\pi(f^2) < \infty$. The space $(L_2(\pi), \|\cdot\|)$, where the norm is induced by the inner product on $L_2(\pi)$, $\langle f, g\rangle = \int \pi(\mathrm{d}x) f(x) g(x)$, is a Hilbert space.

Bounded linear operator

LEMMA 4.1 . Let P be a Markov kernel on $(\mathsf{X}, \mathscr{X})$ with π as an invariant probability measure. Then

$$P : f \mapsto Pf$$

is a bounded linear operator on $L_2(\pi)$. Moreover, $|||P||| = 1$.

Hence, P induces a bounded linear operator on $L_2^0(\pi)$. For notational convenience, we use the same notation P for both the Markov kernel and the operator on $L_2^0(\pi)$.

PROOF. For any $f \in L_2(\pi)$, we have $\pi\left[(Pf)^2\right] \leqslant \pi(P[f^2]) = \pi(f^2)$, which shows that P maps $L_2(\pi)$ into itself. The operator is clearly linear. The previous inequality can also be written as $\|Pf\|^2 \leqslant \|f\|^2$, and therefore $|||P||| \leqslant 1$. This shows that P is a bounded linear operator on $L_2(\pi)$. Since $P1 = 1$, we obtain $\|P1\| = \|1\|$, and hence $|||P||| = 1$. ∎

For the most part, we work with real-valued functions. When studying the spectrum and the resolvent set, we implicitly consider the complexification of $L_2(\pi)$, in which case the inner product is $\langle f, g\rangle = \int \pi(\mathrm{d}x) \bar{f}(x) g(x)$. Moreover, although most of the results in these lecture notes are stated for the Markov operator P, they hold more generally for any bounded linear operator. We focus on Markov operators only to avoid unnecessary generality. We denote by $\mathsf{BL}_2(\pi)$ the set of bounded linear operators on $L_2(\pi)$. We define:

- $\mathrm{Spec}(P) = \{\lambda \in \mathbb{C} : \lambda I - P \text{ is not invertible}\}$, the spectrum of P.
- $\mathrm{Spec}_p(P) = \{\lambda \in \mathbb{C} : \mathrm{Ker}[\lambda I - P] \neq \{0\}\}$, the point spectrum of P.

Clearly, $\mathrm{Spec}_p(P) \subset \mathrm{Spec}(P)$. Moreover, if $S \in \mathsf{BL}_2(\pi)$ with $|||S||| < 1$, then the series $\sum_k S^k$ is normally convergent and can be shown to be the inverse of $I - S$. It follows that for any $\lambda \in \mathbb{C}$ such that $|\lambda| > 1$, $\lambda I - P = \lambda(I - P/\lambda)$ is invertible, and therefore $\mathrm{Spec}(P) \subset \bar{\mathrm{B}}(0, 1)$.

The resolvent set of P is defined as $\mathrm{Res}(P) = \mathrm{Spec}(P)^c$. It is an open set. Indeed, if S is invertible, then writing $T = S(S^{-1}(T-S)+I)$ and taking T sufficiently close to S, we see that T is invertible with inverse $(I+S^{-1}(T-S))^{-1}S^{-1}$.

By definition, an eigenvalue λ of P belongs to the point spectrum; its multiplicity is $\mathrm{Dim}(\lambda I - P)$.

Note that 1 is an eigenvalue of P with multiplicity 1. Indeed, assume that there exists a function $f \in L_2(\pi)$ satisfying $Pf = f$. Then $\pi(f^2) = \pi([Pf]^2) \leqslant \pi(P[f^2]) = \pi(f^2)$, which implies equality in the Cauchy–Schwarz inequality: for π-almost all $x \in \mathsf{X}$, $(Pf(x))^2 = P[f^2](x)$ and hence f is $\pi - a.s.$ constant. (See also the comments in the appendix for this argument.)

We thus obtain the orthogonal decomposition $L_2(\pi) = \mathrm{Span}(1) \overset{\perp}{\oplus} L_2^0(\pi)$, where $L_2^0(\pi) = \{f \in L_2(\pi) : \pi(f) = 0\}$ is a closed subspace, invariant under P.

We are therefore interested in the asymptotic behaviour of

$$\sup_{f \in L_2^0(\pi)} \|P^n f\| = |||P^n|||_{L_2^0(\pi)} = \sup_{h \in L_2(\pi)} \|P^n h - \pi(h)\|.$$

For convenience, we set $H = L_2(\pi)$ and $H_0 = L_2^0(\pi)$.

THEOREM 4.2 . Defining the spectral radius by $\mathrm{Spec.Rad.}(P|_{H_0}) = \{|\lambda| : \lambda \in \mathrm{Spec}(P|_{H_0})\}$, we have

$$\mathrm{Spec.Rad.}(P|_{H_0}) = \lim_n |||P^n|||_{H_0}^{1/n}.$$

PROOF. Let A denote the left-hand side and B the right-hand side. The existence of the limit in the definition of B follows from Fekete's lemma. Indeed, setting $a_n = |||P^n|||$, we have $\log a_{p+q} \leqslant \log a_p + \log a_q$, which implies that $\lim \log a_n/n$ exists and equals $\inf_n \log a_n/n$. This limit may in fact be $-\infty$.

We first show that $A \leqslant B$, which is the easier direction. If we choose $\lambda \in \mathbb{C}$ such that $|\lambda| > B$, then the series $\sum_n (P/\lambda)^n$ converges normally and is the inverse of $I - P/\lambda$, which shows that $\lambda I - P$ is invertible. Hence λ belongs to the resolvent set of P. Therefore $A \leqslant |\lambda|$. Finally, $A \leqslant B$.

We now show $B \leqslant A$. Let us take $\lambda \in \mathbb{C}$ such that $|\lambda| > A$. Then one may define $\phi(z) = (I - zP)^{-1}$ for all $|z| < \lambda^{-1}$. We now prove the Cauchy integral formula. Readers may safely skip this proof on a first reading; it is included for completeness, and because of its elegance and usefulness. For any $r < |\lambda|^{-1}$ and any $z_0 \in \mathbb{C}$ with $|z_0| < r$, define $g(\beta) = \int_0^{2\pi} \frac{\phi(\beta re^{i\theta} + (1-\beta)z_0)}{re^{i\theta} - z_0} re^{i\theta} \, \mathrm{d}\theta$. Since

$$g'(\beta) = \int_0^{2\pi} \phi'(\beta re^{i\theta} + (1-\beta)z_0) \, re^{i\theta} \, \mathrm{d}\theta = \left[\frac{\phi(\beta re^{i\theta} + (1-\beta)z_0)}{i\beta} \right]_0^{2\pi} = 0,$$

we deduce that g is constant and in particular that $g(0) = g(1)$, which can be rewritten as

$$\phi(z_0) \int_0^{2\pi} \frac{1}{1 - (z_0/r)e^{-i\theta}} \, \mathrm{d}\theta = \int_0^{2\pi} \frac{\phi(re^{i\theta})}{1 - (z_0/r)e^{-i\theta}} \, \mathrm{d}\theta.$$

Expanding the integrand, $(1 - (z_0/r)e^{-i\theta})^{-1} = \sum_n (z_0/r)^n e^{-i\theta n}$, and interchanging the series and the integral (which is justified), we obtain, for z_0 in a neighborhood of 0,

$$\phi(z_0) = \sum_{n=0}^{\infty} z_0^n \frac{1}{2\pi} \frac{\int_0^{2\pi} \phi(re^{i\theta}) e^{-in\theta} \, \mathrm{d}\theta}{r^n} = (I - z_0 P)^{-1} = \sum_{n=0}^{\infty} z_0^n P^n.$$

At this point, we may equate the Taylor expansions, which yields

$$P^n = \frac{1}{2\pi} \frac{\int_0^{2\pi} \phi(re^{i\theta}) e^{-in\theta} \, \mathrm{d}\theta}{r^n}$$

for all $n \in \mathbb{N}$. Since ϕ is continuous and therefore bounded on any compact set, there exists a constant C such that $|||P^n||| \leqslant C/r^n$, and hence $\limsup_n |||P^n|||^{1/n} \leqslant 1/r$. As this holds for any $r < |\lambda|^{-1}$, we obtain $B \leqslant |\lambda|$. Finally, $B \leqslant A$, and the proof is complete. ■

Comment on the proof. To be precise, a careful reading of this proof shows that the mapping $\phi : \mathbb{C} \to L_2^0(\pi)$ should be complex differentiable, that is, holomorphic on the ball $B(0, |\lambda|^{-1})$. This property can be verified directly on the resolvent set of P. Indeed, writing

$$I - (z+h)P = (I - zP)\left[(I - zP)^{-1}(-hP) + I\right]$$

we see that, for h sufficiently small, the inverse of $I - (z+h)P$ is given by

$$(I - (z+h)P)^{-1} = \left[(I - zP)^{-1}(-hP) + I\right]^{-1}(I - zP)^{-1} = \sum_{k=0}^{\infty} (-h)^k \left[(I - zP)^{-1}P\right]^k (I - zP)^{-1}.$$

This expansion implies $\phi(z+h) = \phi(z) - h(I - zP)^{-1}P(I - zP)^{-1} + o(|h|)$, showing that ϕ is indeed holomorphic at z.

4.2 Reversibility and self-adjointness.

We now assume that $\boxed{P \text{ is } \pi\text{-reversible}}$, that is, $\pi(\mathrm{d}x)P(x,\mathrm{d}y) = \pi(\mathrm{d}y)P(y,\mathrm{d}x)$. It follows that P is self-adjoint, i.e., $\langle Pf, g\rangle = \langle f, Pg\rangle$. Since P is self-adjoint, $\langle Pf, f\rangle \in \mathbb{R}$.

THEOREM 4.3 . If P is reversible, then

$$\begin{aligned} |||P|||_{L_2^0(\pi)} &= \sup_{\|f\|\leqslant 1, f\in L_2^0(\pi)} \sqrt{\langle Pf, Pf\rangle} = \sup_{\|f\|\leqslant 1, f\in L_2^0(\pi)} |\langle Pf, f\rangle| \\ &= \lim_{n\to\infty} |||P^n|||_{L_2^0(\pi)}^{1/n} = \sup\{|\lambda| : \lambda \in \mathrm{Spec}(P|_{L_2^0(\pi)})\}. \end{aligned}$$

PROOF. Denote the quantities in the preceding equalities by $A = B = C = D = E$. By definition of the triple norm, we clearly have $A = B$. Moreover, $C \leqslant B$ follows from the Cauchy–Schwarz inequality. To show that $B \leqslant C$, write $\langle Pf, \underbrace{Pf/\|Pf\|}_{g}\rangle$ for f of norm 1 and express this quantity in terms of $\langle P(f \pm g), f \pm g\rangle$ using the parallelogram identity (and the fact that P is self-adjoint). We obtain, noting that $\langle Pf, g\rangle = \|Pf\| \in \mathbb{R}$,

$$\|Pf\| = |\langle Pf, g\rangle| = \left|\frac{1}{4}\left[\langle P(f+g), f+g\rangle - \langle P(f-g), f-g\rangle\right]\right| \leqslant \frac{C}{4}\left(\|f+g\|^2 + \|f-g\|^2\right) \leqslant C.$$

Hence $B = C$. Thus $A = B = C$ and $D = E$ (by Theorem 4.2).

Finally, it remains to show that $A = D$. Using the identity $\langle Pf, Pf\rangle = \langle P^2 f, f\rangle$ in the equalities $A = B = C$, we obtain $|||P|||^2 = |||P^2|||$. By induction, this yields $|||P|||^{2^k} = |||P^{2^k}|||$. Therefore, $|||P||| = |||P^{2^k}|||^{1/2^k}$, and letting k tend to infinity, we conclude that $A = D$.

■

The following theorem is stated for the Markov operator P, but we emphasize that it applies more generally to any self-adjoint bounded operator.

THEOREM 4.4 . If P is self-adjoint, then its eigenvalues are real and

$$\mathrm{Spec}(P|_{L_2^0(\pi)}) \subset [m, M],$$

where

$$m = \inf_{f\in L_2^0(\pi),\, \|f\|\leqslant 1} \langle Pf, f\rangle, \qquad M = \sup_{f\in L_2^0(\pi),\, \|f\|\leqslant 1} \langle Pf, f\rangle,$$

and both bounds m and M belong to the spectrum of P.

PROOF. Let $z \notin [m,M]$ and let us show that it belongs to the resolvent set. Let $f \in H_0$. Choose α such that $\langle(\alpha I - P)f, f\rangle = 0$. Then

$$\|(zI-P)f\|^2 = \|(\alpha I - P)f\|^2 + |z-\alpha|^2\|f\|^2 \geqslant |z-\alpha|^2\|f\|^2 \geqslant \Delta\|f\|^2,$$

where we have set $\Delta = d(z,[m,M]) > 0$. This simple inequality shows that z belongs to the resolvent set. Indeed, it successively implies that $\mathrm{Ker}(zI-P) = \{0\}$, that $\mathrm{Ran}(zI-P)$ is closed, and that if $g \in \mathrm{Ran}(zI-P)^\perp$ then $(\bar{z}I - P)g = 0$. Applying the above inequality with $(\bar{z}, g)$ instead of (z, f), we obtain that $g = 0$. Hence $zI - P$ is invertible and, moreover, its inverse is bounded (again by the same inequality). Therefore, z belongs to the resolvent set. This proves the first part of the theorem.

Finally, suppose that $M = |||P|||$. Choose f_n of norm 1 such that $\langle Pf_n, f_n\rangle \to M$. Then

$$\begin{aligned}\|(MI-P)f_n\|^2 &= M^2 + \|Pf_n\|^2 - 2M\langle f_n, Pf_n\rangle \\ &\leqslant 2M^2 - 2M\langle f_n, Pf_n\rangle \to 2M^2 - 2M^2 = 0.\end{aligned}$$

Thus $MI - P$ is not invertible (otherwise, $1 = \|f_n\|^2 \leqslant |||(MI-P)^{-1}|||\,\|(MI-P)f_n\| \to 0$). Therefore $M \in \mathrm{Spec}(P|_{H_0})$.

So far, the proof has been written for P, but the argument also applies to any bounded self-adjoint operator. We can therefore replace P by $Q = MI - P$ to obtain

$$\sup_{f \in L^0_2(\pi),\ \|f\| \leqslant 1} \langle Qf, f\rangle = M - m,$$

$$\inf_{f \in L^0_2(\pi),\ \|f\| \leqslant 1} \langle Qf, f\rangle = 0.$$

It follows that $|||Q||| = M - m$ and consequently (by the previous argument applied with P replaced by $MI - P$) that $M - m \in \mathrm{Spec}((MI-P)|_{H_0})$. This means that

$$(M-m)I - (MI - P) = -(mI - P)$$

is not invertible. Hence we have shown that $m \in \mathrm{Spec}(P)$. If we now suppose that $-m = |||P|||$, we apply the same reasoning by replacing P with $-P$.

■

A careful inspection of the proof shows that

- $\mathrm{Spec}(P|_{L^0_2(\pi)}) \subset \overline{\{\langle f, Pf\rangle \ :\ f \in L^0_2(\pi), \|f\| \leqslant 1\}}$.

4.2.1 Spectral measure

THEOREM 4.5 . If P is self-adjoint, then for any $f \in L^0_2(\pi)$ there exists a finite (nonnegative) measure μ_f supported on $\mathrm{Spec}(P|_{L^0_2(\pi)}) \subset [-1,1]$ such that, for all $n \in \mathbb{N}$,

$$\langle f, P^n f\rangle = \int_{-1}^{1} x^n \mu_f(\mathrm{d}x).$$

Taking $n = 0$ yields $\mu_f([-1,1]) = \pi(f^2) = \mathrm{Var}_\pi(f)$.

The proof of the theorem is omitted; we only sketch the main ideas. We first give a precise meaning to the map $\phi \mapsto \langle f, \phi(P)f\rangle$: it is initially defined for polynomials and then extended to any continuous function ϕ on $\mathrm{Spec}(P|_{L^0_2(\pi)})$ by the density of polynomials, using the Stone–Weierstrass theorem. This

construction yields a nonnegative continuous linear functional on the space of continuous functions on the compact set $\mathrm{Spec}(P|_{L_2^0(\pi)})$, equipped with the supremum norm. The existence of the spectral measure then follows from the Riesz representation theorem.

This theorem allows us to replace P^n by the scalar x^n, greatly simplifying many arguments. This replacement is justified by the spectral theorem. Note that μ_f may charge the points $\{1\}$ or $\{-1\}$.

We define

- $\mathrm{Abs.Spec.Gap}(P) = 1 - \sup\{|\lambda| : \lambda \in \mathrm{Spec}(P|_{H_0})\}$,
- $\mathrm{SpecGap}(P) = 1 - \sup\{\lambda : \lambda \in \mathrm{Spec}(P|_{H_0})\}$.

Moreover, we have the following result.

PROPOSITION 4.6 . Let P be a reversible Markov kernel.

$$\begin{aligned}\mathrm{SpecGap}(P) &= \inf_{f\in H_0, \|f\|\leqslant 1} \langle f, f\rangle - \langle Pf, f\rangle \\ &= \inf_{f\in H_0, \|f\|\leqslant 1} \langle (I-P)f, f\rangle \\ &= \inf_{f\in H_0, \|f\|\leqslant 1} \frac{1}{2}\int \pi(\mathrm{d}x)P(x,\mathrm{d}y)(f(y)-f(x))^2.\end{aligned}$$

Some comments. For the first equality, note that $I-P$ is reversible. Applying Theorem 4.4 with P replaced by $I-P$ gives

$$\mathrm{SpecGap}(P) = \inf\{\lambda \in \mathbb{C} \,:\, \lambda \in \mathrm{Spec}((I-P)|_{H_0})\} = \inf\left\{\langle f, (I-P)f\rangle \,:\, f \in L_2^0(\pi), \|f\| \leqslant 1\right\}$$

The second equality is immediate. The third follows by expanding $\frac{1}{2}\int \pi(\mathrm{d}x)P(x,\mathrm{d}y)(f(y)-f(x))^2$ and using that P is π-invariant. The standard notation for the Dirichlet form is $\mathscr{E}(f,g) = \langle f, (I-P)g\rangle$. We thus have two equivalent expressions for the Dirichlet form $\mathscr{E}(f,f)$:

$$\mathscr{E}(f,f) = \langle f, (I-P)f\rangle = \frac{1}{2}\int \pi(\mathrm{d}x)P(x,\mathrm{d}y)(f(y)-f(x))^2.$$

If P is positive, that is, $\langle f, Pf\rangle \geqslant 0$ for all $f \in L_2^0(\pi)$, then $\mathrm{Spec}(P|_{H_0}) \subset [0,1]$ and the spectral gap coincides with the absolute spectral gap, which allows us to combine Theorem 4.3 and Proposition 4.6.

4.3 Comparison of asymptotic behavior for two Markov kernels

As a consequence of the different expressions for the spectral gap in Proposition 4.6, we have

COROLLARY 4.7 . Let P and Q be π-reversible kernels and assume that $P \succeq Q$ in the sense of covariance ordering, that is, $\langle Pf, f\rangle \leqslant \langle Qf, f\rangle$ for all $f \in H_0$. Then:

- $\mathrm{SpecGap}(P) \geqslant \mathrm{SpecGap}(Q)$.
- $\lim_{n\to\infty} \mathrm{Var}_P\left[\frac{1}{\sqrt{n}}\sum_{k=0}^{n-1} f(X_k)\right] \leqslant \lim_{n\to\infty} \mathrm{Var}_Q\left[\frac{1}{\sqrt{n}}\sum_{k=0}^{n-1} f(X_k)\right]$, where both chains start from the stationary distribution π.

The first bullet follows immediately from covariance ordering: $\langle Pf, f\rangle \leqslant \langle Qf, f\rangle$ is equivalent to $\langle (I-Q)f, f\rangle \leqslant \langle (I-P)f, f\rangle$. If the spectral gaps are positive, then P converges geometrically to π at a faster rate than Q.

The second claim is more delicate; its proof follows Tierney (1998). It shows that the Monte Carlo estimator of $\pi(f)$ has a smaller asymptotic variance when using P rather than Q.

How can covariance ordering be verified? Assume that all singletons are measurable, that is, $\{x\} \in \mathscr{X}$ for every $x \in \mathsf{X}$ (as is the case, for example, when $(\mathsf{X}, \mathscr{X})$ is standard Borel). In many cases, it suffices to show that for all $(x, A) \in \mathsf{X} \times \mathscr{X}$,

$$P(x, A \setminus \{x\}) \geqslant Q(x, A \setminus \{x\}),$$

This condition is known as Peskun ordering. Indeed,

$$\frac{1}{2} \int \pi(\mathrm{d}x) P(x, \mathrm{d}y)(f(y) - f(x))^2 \geqslant \frac{1}{2} \int \pi(\mathrm{d}x) Q(x, \mathrm{d}y)(f(y) - f(x))^2,$$

which implies $\langle (I - P)f, f \rangle \geqslant \langle (I - Q)f, f \rangle$ and hence $P \succeq Q$.

The following exercise enables us to prove the second claim in Corollary 4.7.

EXERCISE 3 . Let P be a reversible Markov kernel and let $f \in H_0$. Define $A_n = \mathrm{Var}_P\left[\frac{1}{\sqrt{n}} \sum_{k=0}^{n-1} f(X_k)\right]$, where the chain starts from the stationary distribution π.

1. Show that
$$A_n = \langle f, f \rangle + 2 \sum_{\ell=1}^{n-1} \frac{n - \ell}{n} \langle f, P^\ell f \rangle$$

2. Deduce that there exists a finite non-negative measure μ_f on $[-1, 1]$ such that
$$A_n = \int_{[-1,1]} w_n(x) \mu_f(\mathrm{d}x) \quad \text{with} \quad w_n(x) = \frac{1 + x}{1 - x} - \frac{2x}{(1 - x)^2} \frac{1 - x^n}{n}$$

3. Split the integral over $[-1, 0]$ and $(0, 1]$ and show that $\lim_{n \to \infty} A_n$ exists and equals
$$-\langle f, f \rangle + 2 \int_{-1}^{1} \frac{1}{1 - x} \mu_f(\mathrm{d}x)$$

We now consider two π-reversible kernels P_0, P_1 such that $P_0 \succeq P_1$ according to the covariance ordering. Define $P_\alpha = (1 - \alpha)P_0 + \alpha P_1$ for $\alpha \in (0, 1]$ and for any $\lambda \in (0, 1)$, write $H_\lambda(\alpha) = (I - \lambda P_\alpha)^{-1}$.

4. Show that H'_λ (the right derivative of H_λ) is equal to
$$H'_\lambda(\alpha) = \lambda (I - \lambda P_\alpha)^{-1} (P_1 - P_0)(I - \lambda P_\alpha)^{-1}.$$

5. Using the fact that P_0, P_1 are π-reversible, show that $\langle f, H'_\lambda(\alpha) f \rangle \geqslant 0$.

6. Deduce $\langle f, H_\lambda(0) f \rangle \leqslant \langle f, H_\lambda(1) f \rangle$.

7. Letting $\lambda \to 1$, deduce that
$$\lim_{n \to \infty} \mathrm{Var}_{P_0}\left[\frac{1}{\sqrt{n}} \sum_{k=0}^{n-1} f(X_k)\right] \leqslant \lim_{n \to \infty} \mathrm{Var}_{P_1}\left[\frac{1}{\sqrt{n}} \sum_{k=0}^{n-1} f(X_k)\right].$$

A Appendix

The following lemma can be used to show that the eigenvalue 1 has multiplicity 1. The proof, however, is more involved than the elementary argument presented in the lecture notes, although it is also more general because it applies to $f \in L_1(\pi)$ rather than just $f \in L_2(\pi)$. That argument was originally due to an MDA student in 2025. We include the proof for $f \in L_1(\pi)$ below for completeness.

LEMMA .8 . If P admits a unique invariant probability measure π, then any harmonic function $f \in L_1(\pi)$ is $\mathbb{P}_\pi - a.s.$ constant.

PROOF. From $Pf = f$, we deduce that $\{f(X_n) \,:\, n \in \mathbb{N}\}$ is a martingale and that $\sup_{n\in\mathbb{N}} \mathbb{E}_\pi[f(X_n)^+] = \pi(f^+) < \infty$, so that it converges $\mathbb{P}_\pi$-almost surely.

We argue by contradiction. If f is not $\mathbb{P}_\pi$-almost surely constant, then there exist $a < b$ such that $\pi(f < a) > 0$ and $\pi(f > b) > 0$. Then, $\mathbb{P}_\pi$-almost surely,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} 1_{\{f(X_k)<a\}} = \pi(f < a) > 0.$$

Hence $\#\{k : f(X_k) < a\} = \infty$, $\mathbb{P}_\pi - a.s.$, and similarly $\#\{k : f(X_k) > b\} = \infty$, $\mathbb{P}_\pi - a.s.$, which contradicts the almost sure convergence of $\{f(X_n) \,:\, n \in \mathbb{N}\}$. ■

Index

approximation lemma, 4
asymptotic variance
    comparison, 25

Birkhoff's theorem, 3
bounded linear operator, 21

central limit theorem, 19
    change of initial distribution, 20
    under drift and minorization, 19
    via Poisson equation, 19
coupling, 9
    inequality, 10
    joint kernel, 14
    maximal, 11
covariance ordering, 25

Dirichlet form, 25
drift condition, 12
drift function, 12
dynamical system, 3

eigenvalue, 22
ergodic dynamical system, 3
ergodicity
    of Markov chains, 3

geometric ergodicity, 13

Hamming distance, 11
harmonic function, 27

invariant
    set, 3
invariant probability measure
    uniqueness and ergodicity, 5

$L_2(\pi)$, 21
law of large numbers
    for almost every initial state, 6
    for arbitrary initial distributions, 7
    Metropolis-Hastings, 7
    under stationarity, 6

Markov operator, 21
martingale decomposition, 18
martingales, 18
    central limit theorem, 18
    increment process, 18
measurable diagonal, 10
Metropolis-Hastings algorithm, 6
    acceptance probability, 6
minorizing condition, 12

Peskun ordering, 26
point spectrum, 21
Poisson equation, 17
    existence of a solution, 17

resolvent set, 22
reversibility, 23

self-adjoint operator, 23
setwise convergence, 7
shift operator, 4
small set, 13
spectral gap, 25
    absolute, 25
spectral measure, 24
spectral radius, 22
spectrum, 21

total variation distance, 10

uniform ergodicity, 13

V-norm, 12