4. Markov Chains: Limit Theorems

All algorithms to be developed have three properties in common: (i) A given configuration is updated in subsequent steps. (ii) Updating in the n^{th} step is performed according to some probabilistic rule. (iii) This rule depends only on the number of the step and on the current configuration. The state of such a system evolves according to some random dynamics which have no memory. Markov chains are appropriate models for such random dynamics (in discrete time).

In this chapter, some abstract limit theorems are derived which later can easily be specialized to prove convergence of various dynamic Monte Carlo methods.

4.1 Preliminaries

The following definitions and remarks address those readers who are not familiar with the basic elements of stochastic processes (with finite state spaces and discrete time). Probabilists will not like this section and those who have met Markov chains should skip it. On the other hand, the author learned in many lectures that students from other fields than mathematics often are grateful for some 'stupid' remarks like those to follow.

We are already acquainted with random transitions, since the observations were random functions of the images. The following definition generalizes this concept.

Definition 4.1.1. *Let X be a finite set called* **state space**. *A family*

$$(P(x,\cdot))_{x\in\mathbf{X}}$$

of probability distributions is called a **transition probability** *or a* **Markov kernel**.

A Markov kernel P can be represented by a matrix – which will be denoted by P as well – where $P(x,y)$ is the element in the x-th row and the y-th coloumn, i.e. a $|\mathbf{X}|\times|\mathbf{X}|$ square matrix with probability vectors in the rows.

If ν is a probability distribution on $\mathbf{X}$ then $\nu(x)P(x,y)$ is the probability to pick x at random from ν and then to pick y at random from $P(x,\cdot)$. The probability of starting anywhere and arriving at y is

$$\nu P(y) = \sum_x \nu(x) P(x,y).$$

Since summation over all y gives 1, νP is a new probability distribution on $\mathbf{X}$. For instance, $\varepsilon_x P(y) = P(x,y)$ for the Dirac distribution ε_x in x (i.e. $\varepsilon(x) = 1$). If we start at x, apply P and then another Markov kernel Q we get y with probability

$$PQ(x,y) = \sum_z P(x,z) Q(z,y).$$

The composition PQ of P and Q is again a Markov kernel as summation over y shows. Note that νP and PQ correspond to multiplication of matrices (if ν is represented by a $1 \times |\mathbf{X}|$ matrix or a row vector). Given ν and kernels P_i one defines recursively $\nu P_1 \dots P_n = (\nu P_1 \dots P_{n-1}) P_n$. All the rules of matrix multiplication apply to the composition of kernels. In particular, composition of kernels is associative.

Definition 4.1.2. *An* **(inhomogeneous) Markov chain** *on the finite space* $\mathbf{X}$ *is given by an initial distribution* ν *and Markov kernels* $P_1, P_2, \dots$ *on* $\mathbf{X}$. *If* $P_i = P$ *for all* i *then the chain is called* **homogeneous**.

Given a Markov chain, the probability that at times $0, \dots, n$ the states are $x_0, x_1, \dots, x_n$ is $\nu(x_0) P_1(x_0, x_1) \dots P_n(x_{n-1}, x_n)$. This defines a probability distribution $\mathsf{P}^{(n)}$ on the space $\mathbf{X}^{\{0,\dots,n\}}$ of such sequences of length $n+1$. These distributions are consistent, i.e. $\mathsf{P}^{(n+1)}$ induces $\mathsf{P}^{(n)}$ by

$$\mathsf{P}^{(n)}((x_0, \dots, x_n)) = \sum_{x_{n+1}} \mathsf{P}^{(n+1)}((x_0, \dots, x_n, x_{n+1})).$$

An infinite sequence $(x_0, \dots x_n, \dots)$ of states is called a **path** (of the Markov chain). The set of all paths is $\mathbf{X}^{\mathbb{N}_0}$. Because of consistency, one can define the probability of those sets of paths, which depend on a finite number of time indices only: Let $A \subset \mathbf{X}^{\mathbb{N}_0}$ be a (finite cylinder) set $A = B \times \mathbf{X}^{\{n+1,\dots\}}$ with $B \subset \mathbf{X}^{\{0,\dots,n\}}$. Then $\mathsf{P}(A) = \mathsf{P}^{(n)}(B)$ is called the probability of A (w.r.t. the given chain).

Remark 4.1.1. The concept of probability was extended from the subsets of a finite set to a class of subsets of an infinite space. It does not contain sets of paths which depend on an infinite number of times, for example defined by a property like 'the path visits state 1 infinitely often'. For applications using such sets the above concept is too narrow. The extension to a probability distribution on a sufficiently large class of sets involves some measure theory. It can be found in almost any introduction to probability theory above the elementary level (e.g. BILLINGSLEY (1979)). For the development of the algorithms in the next chapters this extension is not necessary. It will be needed only for some more advanced considerations in later chapters.

Markov chains can also be introduced via sequences of random variables ξ_i fulfilling the **Markov property**

$$\mathsf{P}(\xi_n = x_n | \xi_0 = x_0, \dots, \xi_{n-1} = x_{n-1}) = \mathsf{P}(\xi_n = x_n | \xi_{n-1} = x_{n-1})$$

for all $n \geq 1$ and $x_0, \dots, x_n \in \mathbf{X}$. To obtain Markov chains in the above sense let $\nu(x) = \mathsf{P}(\xi_0 = x)$ be the initial distribution and let the transition probabilities be given by the conditional probablities

$$P_n(x,y) = \mathsf{P}(\xi_n = y | \xi_{n-1} = x).$$

Conversely, given the initial distribution and the transition probabilities the random variables ξ_i can be defined as the projections of $\mathbf{X}^{\mathbb{N}_0}$ onto the coordinates, i.e. the maps

$$\xi_n : \mathbf{X}^{\{0,\dots\}} \longrightarrow \mathbf{X}, (x_i)_{i \geq 0} \longmapsto x_n.$$

Example 4.1.1. Let us compute the probabilities of some special events for the chain (ξ_i) of projections. In the following computations all denominators are assumed to be strictly positive.

(a) The distribution of the chain in the n-th step is

$$\begin{aligned} \nu_n(x) &= \mathsf{P}(\xi_n = x) = \sum_{x_0,\dots,x_{n-1}} \mathsf{P}((x_0, \dots, x_{n-1}, x)) \\ &= \sum_{x_0,\dots,x_{n-1}} \nu(x_0) P_1(x_0, x_1) \cdot \dots \cdot P_n(x_{n-1}, x) = \nu P_1 \dots P_n(x). \end{aligned}$$

ν_n is called (the n-th) **one-dimensional marginal distribution** of the process.

(b) For $m < n$, the **two-dimensional marginals** are given by

$$\begin{aligned} &\nu_{mn}(x,y) = \mathsf{P}(\xi_m = x, \xi_n = y) \\ &= \sum_{x_0,\dots,x_{m-1}} \sum_{x_{m+1},\dots x_{n-1}} \mathsf{P}((x_0, \dots, x_{m-1}, x, x_{m+1}, \dots, x_{n-1}, y)) \\ &= \nu P_1 \dots P_m(x) P_{m+1} \dots P_n(x,y). \end{aligned}$$

(c) Defining a Markov process via transition probabilities and via the projections ξ_i is consistent:

$$\begin{aligned} \mathsf{P}(\xi_n = y | \xi_{n-1} = x) &= \frac{\mathsf{P}(\xi_{n-1} = x, \xi_n = y)}{\mathsf{P}(\xi_{n-1} = x)} = \frac{\nu_{n-1,n}(x,y)}{\nu_{n-1}(x)} \\ &= \frac{\nu P_1 \dots P_{n-1}(x) P_n(x,y)}{\nu P_1 \dots P_{n-1}(x)} = P_n(x,y). \end{aligned}$$

It is now easy to check the Markov property of the projections:

$$\begin{aligned}
&\mathsf{P}(\xi_n = y | \xi_0 = x_0, \dots, \xi_{n-1} = x) \\
&= \frac{\mathsf{P}(\xi_0 = x_0, \dots, \xi_{n-1} = x, \xi_n = y)}{\sum_z \mathsf{P}(\xi_0 = x_0, \dots, \xi_{n-1} = x, \xi_n = z)} \\
&= \frac{\nu(x_0) P_1(x_0, x_1) \dots P_{n-1}(x_{n-1}, x) P_n(x, y)}{\sum_z \nu(x_0) P_1(x_0, x_1) \dots P_{n-1}(x_{n-1}, x) P_n(x, z)} \\
&= P_n(x, y) = \mathsf{P}(\xi_n = y | \xi_{n-1} = x).
\end{aligned}$$

Expressions like those in (a) and (b) can be derived also for the higher dimensional marginal distributions $\mathsf{P}(\xi_{n_1} = x_1, \dots, \xi_{n_k} = x_k)$. We shall sometimes call P the **law** of the Markov chain $(\xi_n)_{n \geq 0}$. Given P the expectation $\mathsf{E}(f)$ is defined in the usual way for those functions f on $\mathbf{X}^{\mathbb{N}_0}$ which depend on a finite number of time indices only. More precisely, if there is $k \geq 0$ such that for all $(x_n)_{n \geq 0}$, $f((x_n)_{n \geq 0}) = f(x_1, \dots, x_k)$ then

$$\mathsf{E}(f) = \sum_{x_0, \dots, x_k} f(x_0, \dots, x_k) \mathsf{P}((x_0, \dots, x_k)).$$

Example 4.1.2. Let $x \in \mathbf{X}$ be fixed. Then $h((x_i)_{i \geq 0}) = \sum_{i=0}^n 1_{\{\xi_i = x\}}$ is the number of visits of the path (x_i) in x up to time n. The expected number of visits is

$$\mathsf{E}(h) = \sum_{y_0, \dots, y_n} h(y_0, \dots, y_n) \mathsf{P}((y_0, \dots, y_n)) = \sum_{i=0}^n \nu_i(x).$$

We will be interested in the limiting behaviour of Markov chains. Two concepts of convergence will be used: Let ξ and $\xi_0, \xi_1, \dots$ be random variables. We shall say that (ξ_i) converges to ξ

(a) **in probability** if for every $\varepsilon > 0$,

$$\mathsf{P}(|\xi_i - \xi| > \varepsilon) \longrightarrow 0, \quad i \to \infty;$$

(b) **in** L^2, if

$$\mathsf{E}((\xi_i - \xi)^2) \longrightarrow 0, \quad i \to \infty.$$

For every nonnegative random variable η, Markov's inequality states that

$$\mathsf{P}(\eta \geq \varepsilon) \leq \frac{\mathsf{E}(\eta^2)}{\varepsilon^2}.$$

By this inequality,

$$\mathsf{P}(|\xi_i - \xi| > \varepsilon) \leq \frac{\mathsf{E}((\xi_i - \xi)^2)}{\varepsilon^2}$$

and hence L^2 -convergence implies convergence in probability. For bounded functions the two concepts are equivalent.

Let us finally note that a Markov chain with strictly positive initial distribution and transition probabilities induces a (finite) Markov field in a natural

way: on each time interval $I = \{0, \dots, n\}$ define a neighbourhood system by $\partial(k) = \{k-1, k+1\} \cap I$. Then for $k \in I \backslash \{0\}$,

$$\begin{aligned}
\mathsf{P}(\xi_k | \xi_i, 0 \leq i \leq n, i \neq k) &= \frac{\mathsf{P}(\xi_i = x_i, 0 \leq i \leq n)}{\mathsf{P}(\xi_i, 0 \leq i \leq n, i \neq k)} \\
&= \frac{\nu(x_0) P_1(x_0, x_1) \dots P_{k-1}(x_{k-2}, x_{k-1}) P_k(x_{k-1}, x_k)}{\sum_z \nu(x_0) P_1(x_0, x_1) \dots P_{k-1}(x_{k-2}, x_{k-1}) P_k(x_{k-1}, z)} \\
&\quad \frac{P_{k+1}(x_k, x_{k+1}) \dots P_n(x_{n-1}, x_n)}{P_{k+1}(z, x_{k+1}) \dots P_n(x_{n-1}, x_n)} \\
&= \frac{\nu_{k-1}(x_{k-1}) P_k(x_{k-1}, x_k) P_{k+1}(x_k, x_{k+1})}{\sum_z \nu_{k-1}(x_{k-1}) P_k(x_{k-1}, z) P_{k+1}(z, x_{k+1})} \\
&= \frac{\mathsf{P}(\xi_{k-1} = x_{k-1}, \xi_k = x_k, \xi_{k+1} = x_{k+1})}{\mathsf{P}(\xi_{k-1} = x_{k-1}, \xi_{k+1} = x_{k+1})} \\
&= \mathsf{P}(\xi_k = x_k | \xi_{k-1} = x_{k-1}, \xi_{k+1} = x_{k+1})
\end{aligned}$$

and similarly,

$$\mathsf{P}(\xi_0 = x_0 | \xi_i = x_i, 1 \leq i \leq n) = \mathsf{P}(\xi_0 = x_0 | \xi_1 = x_1).$$

This is the spatial Markov property we met in Chapter 3.

Markov chains are introduced at an elementary level in KEMENEY and SNELL (1960). Those who prefer a more formal (matrix-theoretic) treatment may consult SENETA (1981).

4.2 The Contraction Coefficient

To prove the basic limit theorems for homogeneous and inhomogeneous Markov chains, the classical contraction method is adopted, a remarkably simple and transparent argument. The proofs are given explicitely for finite state spaces. Adopting the proper definition of total variation and replacing some of the 'max' by 'l.u.b.' essentially yields the corresponding results for more general spaces.

The special structure of the configuration space $\mathbf{X}$ presently is not needed. Hence $\mathbf{X}$ is merely assumed to be a finite set.

For distributions μ and ν on $\mathbf{X}$, the norm of **total variation** of the difference $\mu - \nu$ is given by

$$\|\mu - \nu\| = \sum_x |\mu(x) - \nu(x)|.$$

Note that this simply is the L^1-norm of the difference. The following equivalent descriptions are useful.

Lemma 4.2.1. *Let μ and ν be probability distributions on $\mathbf{X}$. Then*

$$\begin{aligned}\|\mu-\nu\| &= 2\sum_x(\mu(x)-\nu(x))^+\\ &= 2(1-\sum_x \mu(x)\wedge\nu(x))\\ &= \max\left\{\left|\sum_x h(x)(\mu(x)-\nu(x))\right| : |h|\le 1\right\}.\end{aligned}$$

For a vector $\rho=(\rho(x))_{x\in\mathbf{X}}$ the positive part ρ^+ equals $\rho(x)$ if $\rho(x)>0$ and vanishes otherwise. The negative part ρ^- is $(-\rho)^+$. The symbol $a\wedge b$ denotes the minimum of real numbers a and b.

If $\mathbf{X}$ is not finite a definition of total variation is obtained replacing the sum in the last expression by the integral $\int h\,d(\mu-\nu)$ and the maximum by the least upper bound.

Remark 4.2.1. For probability distributions μ and ν the triangle inequality yields $\|\mu-\nu\|\le 2$. From the second identity in the lemma one reads off that equality holds if and only if μ and ν have disjoint support (the **support** of a distribution ν is the set where it is strictly positive; two distributions with disjoint support are called **orthogonal**).

Proof (of Lemma 4.2.1). Plainly,

$$\begin{aligned}\|\mu-\nu\| &= \sum_x(\mu(x)-\nu(x))^+ + \sum_x(\mu(x)-\nu(x))^-\\ &= \sum_{x:\mu(x)\ge\nu(x)}(\mu(x)-\nu(x)) + \sum_{x:\mu(x)<\nu(x)}(\nu(x)-\mu(x)).\end{aligned}$$

The difference of the sums vanishes since μ and ν are probability distributions and hence the sums are equal. This yields

$$\|\mu-\nu\|/2 = \sum_x(\mu(x)-\nu(x))^+$$

and hence the first identity. Furthermore,

$$\begin{aligned}\|\mu-\nu\|/2 &= \sum_{x:\mu(x)\ge\nu(x)}\mu(x) - \sum_{x:\mu(x)\ge\nu(x)}\nu(x)\\ &= \sum_x\mu(x) - \sum_{x:\mu(x)<\nu(x)}\mu(x) - \sum_{x:\mu(x)\ge\nu(x)}\nu(x)\\ &= 1-\sum_x\mu(x)\wedge\nu(x)\end{aligned}$$

which proves the second identity. Finally, the inequality

$$\begin{aligned}\|\mu-\nu\| &= \sum_x|\mu(x)-\nu(x)|\\ &\ge \max\left\{\left|\sum_x h(x)(\mu(x)-\nu(x))\right| : |h|\le 1\right\}\end{aligned}$$

is obvious. To check equality plug in $h(x)=\operatorname{sgn}(\mu(x)-\nu(x))$. □

The **contraction coefficient** of a Markov kernel P is defined by

$$c(P)=(1/2)\max_{x,y}\|P(x,\cdot)-P(y,\cdot)\|.$$

The notion of a contraction coefficient can be considerably generalized, cf. SENETA (1981), 4.3.

Remark 4.2.2. By the last remark, $c(P)\le 1$ and equality holds if and only if at least two of the distributions $P(x,\cdot)$ have disjoint support. Plainly, $c(P)=0$ if and only if all $P(x,\cdot)$ are equal. Hence the contraction coefficient is a rough measure for orthogonality of the distributions $P(x,\cdot)$.

The name 'contraction coefficient' is justified by the next inequality. This and the following one are nearly all what is needed to prove the ergodic theorems below.

Lemma 4.2.2. *Let μ and ν be probability distributions and P and Q be Markov kernels on $\mathbf{X}$. Then*

$$\begin{aligned}\|\mu P-\nu P\| &\le c(P)\|\mu-\nu\|,\\ c(PQ) &\le c(P)c(Q).\end{aligned}$$

In particular,

$$\begin{aligned}\|\mu P-\nu P\| &\le \|\mu-\nu\|,\\ \|\mu P-\nu P\| &\le 2c(P).\end{aligned}$$

Proof. Let us start with the first inequality. For a real function f on $\mathbf{X}$ let

$$d=(\max_x f(x)+\min_x f(x))/2.$$

Then

$$\max_x|f(x)-d|=(1/2)\max_{x,y}|f(x)-f(y)|.$$

Writing $\mu(f)$ for $\sum_x f(x)\mu(x)$, we conclude

$$\begin{aligned}|\mu(f)-\nu(f)| &= |\mu(f-d)-\nu(f-d)|\le\max_x|f(x)-d|\cdot\|\mu-\nu\|\\ &= (1/2)\max_{x,y}|f(x)-f(y)|\cdot\|\mu-\nu\|.\end{aligned}\tag{4.1}$$

For a function h on $\mathbf{X}$, the function Ph is defined by

$$Ph(x) = \sum_y h(y)P(x,y).$$

Plugging in Ph for f yields

$$\begin{aligned}
\|\mu P - \nu P\| &= \max\{|(\mu P)h - (\nu P)h| : |h| \le 1\} \\
&= \max\{|\mu(Ph) - \nu(Ph)| : |h| \le 1\} \\
&\le \max\left\{(1/2)\max_{x,y}|Ph(x) - Ph(y)| : |h| \le 1\right\}\|\mu - \nu\| \\
&= (1/2)\max_{x,y}\max\{|Ph(x) - Ph(y)| : |h| \le 1\}\|\mu - \nu\| \\
&= c(P)\|\mu - \nu\|
\end{aligned}$$

and hence the first inequality. The second one follows from

$$\begin{aligned}
c(PQ) &= (1/2)\max_{x,y}\|PQ(x,\cdot) - PQ(y,\cdot)\| \\
&= (1/2)\max_{x,y}\|P(x,\cdot)Q - P(y,\cdot)Q\| \\
&\le c(P)c(Q).
\end{aligned}$$

The other inequalities follow from the first two since $c(P) \le 1$ and $\|\mu - \nu\| \le 2$. This completes the proof. □

Remark 4.2.3. An immediate consequence is **asymptotic loss of memory** or **weak ergodicity** of Markov chains:

Let P_n, $n \ge 1$, be Markov kernels and μ and ν two initial distributions. Then $c(P_1 \dots P_n) \to 0$ implies

$$\|\mu P_1 \dots P_n - \nu P_1 \dots P_n\| \longrightarrow 0.$$

Markov chains will converge quickly if the contraction coefficient is small. Therefore the following estimate is useful.

Lemma 4.2.3. *For every Markov kernel Q on a finite space $\mathbf{X}$,*

$$c(Q) \le 1 - |\mathbf{X}|\min\{Q(x,y) : x,y \in \mathbf{X}\} \le 1 - \min\{Q(x,y) : x,y \in \mathbf{X}\}.$$

In particular, if Q is strictly positive then $c(Q) < 1$.

Proof. By Lemma 4.2.1,

$$\|\mu - \nu\|/2 = 1 - \sum_x \mu(x) \wedge \nu(x)$$

for probability distributions μ and ν. Hence

$$c(Q) = 1 - \min\left\{\sum_z Q(x,z) \wedge Q(y,z) : x,y \in \mathbf{X}\right\}$$

which implies the first two inequalities. The rest is an immediate consequence. □

4.3 Homogeneous Markov Chains

A Markov chain is called **homogeneous** if all its transition probabilities are equal. We prove convergence of marginals and a law of large numbers for homogeneous Markov chains.

Lemma 4.3.1. *For each Markov kernel P on a finite state space the sequence $(c(P^n))_{n\ge 0}$ decreases. If P has a strictly positive power P^τ then the sequence decreases to 0.*

Markov kernels with a strictly positive power are called **primitive**. A homogeneous chain with primitive Markov kernel eventually reaches each state with positive probability from any state. This property is called **irreducibility** (a characterization of primitive Markov kernels more common in probability theory is to say that they are irreducible and aperiodic, cf. SENETA (1981)).

Proof (of Lemma 4.3.1). By Lemma 4.2.2,

$$c(P^{n+1}) \le c(P)c(P^n) \le c(P^n).$$

If $Q = P^\tau$ then

$$c(P^n) \le (Q^k P^{n-\tau k}) \le c(Q)^k$$

for $n \ge \tau$ and the greatest number k with $\tau k \le n$. If Q is strictly positive then $c(Q) < 1$ by Lemma 4.2.3 and $c(P^n)$ tends to zero as n tends to infinity. This proves the assertion. □

Let μ be a probability distribution on $\mathbf{X}$. If $\mu P = \mu$ then $\mu P^n = \mu$ for every $n \ge 0$ and hence such distributions are natural candidates for limit distributions of homogeneous Markov chains. A distribution μ satisfying $\mu P = \mu$ is called **invariant** or **stationary** for P. The limit theorem reads:

Theorem 4.3.1. *A primitive Markov kernel P on a finite space has a unique invariant distribution μ and*

$$\nu P^n \longrightarrow \mu \quad as \quad n \to \infty$$

uniformly in all distributions ν.

Proof. Existence and uniqueness of the invariant distribution is part of the Perron-Frobenius theorem (Appendix B). By Lemma 4.3.1, the sequence $(c(P^n))$ decreases to zero and the theorem follows from

$$\|\nu P^n - \mu\| = \|\nu P^n - \mu P^n\| \le \|\nu - \mu\|c(P^n) \le 2 \cdot c(P^n). \qquad (4.2)$$

□

Homogeneous Markov chains with primitive kernel even obey the law of large numbers. For an initial distribution ν and a Markov kernel P let $(\xi_i)_{i\geq 0}$ be a corresponding sequence of random variables (cf. Section 4.1). The expectation $\sum_x f(x)\mu(x)$ of a function f on $\mathbf{X}$ w.r.t. a distribution μ will be denoted by $\mathsf{E}_\mu(f)$.

Theorem 4.3.2 (Law of Large Numbers). *Let $\mathbf{X}$ be a finite space and let P be a primitive Markov kernel on $\mathbf{X}$ with invariant distribution μ. Then for every initial distribution ν and every function f on $\mathbf{X}$,*

$$\frac{1}{n}\sum_{i=1}^{n} f(\xi_i) \longrightarrow \mathsf{E}_\mu(f)$$

in $L^2(\mathsf{P}_\nu)$. Moreover, for every $\varepsilon > 0$,

$$\mathsf{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} f(\xi_i) - \mathsf{E}_\mu(f)\right| > \varepsilon\right) \leq \frac{13\|f\|^2}{(1-c(P))n\varepsilon^2}$$

where $\|f\| = \sum_x |f(x)|$.

For identically distributed independent random variables ξ_i the Markov kernel $(P(x,y))$ does not depend on x, hence the rows of the matrix coincide and $c(P) = 0$. In this case the theorem boils down to the usual weak law of large numbers.

Proof. Choose $x \in \mathbf{X}$ and let $f = 1_{\{x\}}$. By elementary calculations,

$$\begin{aligned}
&\mathsf{E}\left(\left(\frac{1}{n}\sum_{i=1}^{n} f(\xi_i) - \mathsf{E}_\mu(f)\right)^2\right)\\
=\ &\mathsf{E}\left(\left(\frac{1}{n}\sum_{i=1}^{n} 1_{\{\xi_i=x\}} - \mu(x)\right)^2\right)\\
=\ &\frac{1}{n^2}\sum_{i,j=1}^{n} \mathsf{E}((1_{\{\xi_i=x\}} - \mu(x))(1_{\{\xi_j=x\}} - \mu(x)))\\
=\ &\frac{1}{n^2}\sum_{i,j=1}^{n} ((\nu_{ij}(x,x) - \mu(x)^2) - (\mu(x)\nu_j(x) - \mu(x)^2)\\
&-(\mu(x)\nu_i(x) - \mu(x)^2)).
\end{aligned}$$

There are three means to be estimated. The first one is most difficult. Since $\mu P = \mu$, for $i, k > 0$ and $x, y \in \mathbf{X}$ the following rough estimates hold:

$$\begin{aligned}
&|\nu P^i(x)\varepsilon_x P^k(y) - \mu(x)\mu(y)|\\
\leq\ &|\nu P^i(x)\varepsilon_x P^k(y) - \mu P^i(x)\varepsilon_x P^k(y)|\\
&+|\mu(x)\varepsilon_x P^k(y) - \mu(x)\mu P^k(y)|\\
\leq\ &\|(\nu-\mu)P^i\| + \|(\varepsilon_x - \mu)P^k\|\\
\leq\ &2\cdot(c(P)^i + c(P)^k).
\end{aligned}$$

Using the explicit expression

$$\sum_{i=1}^{n} a^i = a\cdot\frac{1-a^n}{1-a}, \quad 0 < a < 1,$$

for the finite geometric series, one computes

$$\begin{aligned}
&\frac{1}{n^2}\sum_{i=1}^{n-1}\sum_{j>i}^{n} |\nu_{ij}(x,y) - \mu(x)\mu(y)|\\
\leq\ &2\left(\frac{c(P)}{1-c(P)}\frac{1}{n}(1-c(P)^{n-1}) + \frac{c(P)}{1-c(P)}\frac{n-1}{n^2}(1-c(P)^n)\right)\\
\leq\ &\frac{4}{1-c(P)}\frac{1}{n}.
\end{aligned}$$

The same estimate holds for the mean over pairs (i,j) of indices with $j < i$. For convenience of notation set $\nu_{ii}(x,x) = \nu P^i(x)$ and $\nu_{ii}(x,y) = 0$ if $x \neq y$. The sum over the corresponding terms is bounded by n and hence

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} |\nu_{ij}(x,y) - \mu(x)\mu(y)| \leq \frac{9}{1-c(P)}\frac{1}{n}.$$

By (4.2) the second and third mean can be estimated:

$$\begin{aligned}
&\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} |\mu(x)\nu_j(y) - \mu(x)\mu(y)|\\
=\ &\mu(x)\frac{c(P)}{1-c(P)}\frac{2}{n}(1-c(P)^n) \leq \frac{2}{1-c(P)}\frac{1}{n}.
\end{aligned}$$

Hence the above expectation is bounded by $(13/n)(1-c(P))^{-1}$. For general f, the triangle inequality gives a bound $(c/n)(1-c(P))^{-1}$ with $c = 13\|f\|^2$, $\|f\| = \sum_x |f(x)|$. This proves the first part of the theorem. The second one follows form Markov's inequality. □

Remark 4.3.1. [continuous state space] With a little extra work the above program (and also the extension to inhomogeneous chains in the next section) can be carried out on abstract measurable spaces. Madson and Isaakson (1973) give proofs for the special case

$$P(x, dy) = f_x(y)\, d\nu(x)$$

with densities f_x w.r.t. a σ-finite measure ν. In particular, they cover the important case of densities w.r.t. Lebesgue measure on $\mathbf{X} = \mathsf{R}^d$. They also indicate the extension to the case where densities do not exist. This type of extension is carried out in M. IOSIFESCU (1972). Some remarks on the limits of the contraction technique can be found in Remark 5.1.2.

4.4 Inhomogeneous Markov Chains

Let us now turn to inhomogeneous Markov chains. We first note a simple observation.

Lemma 4.4.1. *If μ_n, $n \geq 1$, are probability distributions on $\mathbf{X}$ such that $\sum_n \|\mu_{n+1} - \mu_n\| < \infty$ then there is a probability distribution μ_∞ such that $\mu_n \to \mu_\infty$ (in $\|\cdot\|$) as $n \to \infty$.*

Since $\mathbf{X}$ is finite, pointwise convergence and convergence in the L^1-norm $\|\cdot\|$ coincide.

Proof. For $m < n$,

$$\|\mu_n - \mu_m\| \leq \sum_{k \geq m} \|\mu_{k+1} - \mu_k\|$$

which tends to zero as m tends to infinity. Thus (μ_n) is a Cauchy sequence in the compact space $\{\mu \in \mathsf{R}^{\mathbf{X}} : \mu \geq 0, \sum_x \mu(x) = 1\}$ and hence has a limit μ_∞ in this set. □

The limit theorem for inhomogeneous Markov chains reads:

Theorem 4.4.1. *Let P_n, $n \geq 1$, be Markov kernels and assume that each P_n has an invariant probability distribution μ_n. Assume further that the following conditions are satisfied*

$$\sum_n \|\mu_n - \mu_{n+1}\| < \infty, \tag{4.3}$$

$$\lim_{n \to \infty} c(P_i \dots P_n) = 0 \quad \text{for every} \quad i \geq 1. \tag{4.4}$$

Then $\mu_\infty = \lim_{n \to \infty} \mu_n$ exists and uniformly in all initial distributions ν,

$$\nu P_1 \dots P_n \longrightarrow \mu_\infty \quad \text{for} \quad n \to \infty.$$

Proof. The existence of the limit μ_∞ was proved in the preceding lemma. Let now $i \geq 1$ and $k \geq 1$. Use $\mu_n P_n = \mu_n$ for

$$\begin{aligned} & \mu_\infty P_i \dots P_{i+k} - \mu_\infty \\ = \; & (\mu_\infty - \mu_i) P_i \dots P_{i+k} + \mu_i P_{i+1} \dots P_{i+k} - \mu_\infty \\ = \; & (\mu_\infty - \mu_i) P_i \dots P_{i+k} + \sum_{j=1}^{k} (\mu_{i-1+j} - \mu_{i+j}) P_{i+j} \dots P_{i+k} \\ & + \mu_{i+k} - \mu_\infty. \end{aligned}$$

For $i \geq N$ this implies

$$\|\mu_\infty P_i \dots P_{i+k} - \mu_\infty\| \leq 2 \cdot \sup_{n \geq N} \|\mu_\infty - \mu_n\| + \sum_{n \geq N} \|\mu_n - \mu_{n+1}\|. \tag{4.5}$$

We used Lemma 4.2.2 and that the contraction coefficient is bounded by 1. By condition (4.3) and since μ_∞ exists, for large N the expression on the right hand becomes small. Fix now a large N. For $2 \leq N \leq i \leq n$ we may continue with

$$\begin{aligned} & \|\nu P_1 \dots P_n - \mu_\infty\| \\ = \; & \|(\nu P_1 \dots P_{i-1} - \mu_\infty) P_i \dots P_n + \mu_\infty P_i \dots P_n - \mu_\infty\| \qquad (4.6) \\ \leq \; & 2 \cdot c(P_i \dots P_n) + \|\mu_\infty P_i \dots P_n - \mu_\infty\| \end{aligned}$$

For large n, the first term becomes small by (4.4). This proves the result. □

The proof shows that convergence of inhomogeneous chains basically is asymptotic loss of memory plus convergence of the invariant distributions.

The theorem frequently is referred to as DOBRUSHIN's theorem (DOBRUSHIN (1956). There are various closely related approaches and it can even be traced back to MARKOV (cf. SENETA (1973) and (1981), pp. 144-145). The contraction technique is exploited systematically in ISAACSON and MADSON (1976).

There are some simple but useful criteria for the conditions in the theorem.

Lemma 4.4.2. *For probability distributions μ_n, $n \geq 1$, condition (4.3) is fulfilled if each of the sequences $(\mu_n(x))_{n \geq 1}$ de- or increases eventually.*

Proof. By Lemma 4.2.1,

$$0 \leq \sum_n \|\mu_{n+1} - \mu_n\| = 2 \sum_x \sum_n (\mu_{n+1}(x) - \mu_n(x))^+.$$

By monotony, there is n_0 such that either $(\mu_{n+1}(x) - \mu_n(x))^+ = 0$ for all $n \geq n_0$ and thus $\sum_{n \geq n_0} (\mu_{n+1}(x) - \mu_n(x))^+ = 0$ or $(\mu_{n+1}(x) - \mu_n(x))^+ = \mu_{n+1}(x) - \mu_n(x)$, and thus

$$\sum_{n=n_0}^{N} (\mu_{n+1}(x) - \mu_n(x))^+ = \mu_{N+1}(x) - \mu_{n_0}(x) \leq 1$$

for all large N. This implies that the double sum is finite and hence condition (4.3) holds. □

Lemma 4.4.3. *Condition (4.4) is implied by*

$$\prod_{k\geq i} c(P_k) = 0 \quad \textit{for every} \quad i \geq 1. \tag{4.7}$$

or by

$$c(P_n) > 0 \quad \textit{for every } n \textit{ and} \quad \prod_{k\geq 1} c(P_k) = 0. \tag{4.8}$$

Proof. Condition (4.7) implies (4.4) by the second rule in Lemma 4.2.2 and obviously (4.8) implies (4.7). □

This can be used to check convergence of a given inhomogeneous Markov chain in the following way: The time axis is subdivided into 'epochs' $(\tau(k-1), \tau(k)]$ over which the transitions

$$Q_k = P_{\tau(k-1)+1} \dots P_{\tau(k)}$$

are strictly positive (and hence also the minimum in the above estimate). Given a time i and a large n there are some epochs inbetween and

$$\begin{aligned} & c(P_i \dots P_n) \\ \leq\ & c(P_i \dots P_{\tau(p-1)}) c(Q_p \dots Q_r) c(P_{\tau(r)+1} \dots P_n) \\ \leq\ & c(Q_p) \dots c(Q_r) \\ \leq\ & \prod_{k=p}^{r} \left(1 - |\mathbf{X}| \min_{x,y} Q_k(x,y)\right). \end{aligned}$$

In order to ensure convergence, the factors (which are strictly smaller than 1) have to be small enough to let the product converge to zero, i.e. the numbers $\min_{x,y} Q_k(x,y)$ should not decrease too fast.

The following comments concern condition (4.4).

Example 4.4.1. It is easy to see that condition (4.4) cannot be dropped: for each n let $P_n = I$ where I is the unit matrix. Then $c(P_n) = 1$, every probability disribution ρ is invariant w.r.t. P_n and (4.3) holds for $\mu_n = \rho$. On the other hand $\nu P_1 \dots P_n \to \nu$ for every ν. One can modify this example such that the μ_n are the *unique* invariant distributions for the P_n. Let

$$P_n = \begin{pmatrix} 1-a_n & a_n \\ a_n & 1-a_n \end{pmatrix}$$

with small positive numbers a_n. For these Markov kernels the uniform distribution $\mu = (1/2, 1/2)$ is the unique invariant distribution. The contraction coefficients are $c(P_n) = |1 - 2a_n|$. There are a_n such that

$$\prod_{n\geq 1} c(P_n) = \prod_{n\geq 1} (1 - 2a_n) \geq \frac{3}{4}.$$

(or which amounts to the same $\sum_n \ln(1-2a_n) \geq \ln(3/4)$). Let now $\nu = (1,0)$ be the initial distribution. Then the one-dimensional marginals $\nu_n = (\nu_n(1), \nu_n(2)) = \nu P_1 \dots P_n$ of the chain fulfill

$$\nu_n(1) \geq (1-a_1)(1-a_2)\dots(1-a_n) \geq (3/4) \quad \text{for each} \quad n$$

and hence do not converge to μ.

Similarly, conditions (4.4), (4.7) or (4.8) cannot be replaced by

$$c(P_1 \dots P_n) \to 0$$

or

$$\prod_k c(P_k) = 0,$$

respectively. In the example, $\nu_1 = (1-a_1, a_1)$. If P_1 is replaced by

$$P_1 = \begin{pmatrix} 1-a_1 & a_1 \\ 1-a_1 & a_1 \end{pmatrix},$$

then $\nu P_1 = (1-a_1, a_1)$ for every initial distribution ν. Convergence of this chain is the same as before but $\prod_k c(P_k) = 0$ since $c(P_1) = 0$.

The Remarks 4.3.1 on continuous state spaces hold for inhomogeneous chains as well.