

Chapter 2

Markov Chains

2.1 The fundamental Markov property

Definition 2.1 A sequence X_0, X_1, X_2, \dots of stochastic variables with values in a common space $(\mathcal{X}, \mathbb{E})$ is a **Markov chain** if

$$X_{n+1} \perp\!\!\!\perp (X_0, X_1, \dots, X_{n-1}) \mid X_n \quad \text{for } n = 1, 2, \dots \quad (2.1)$$

We refer to (2.1) as the **fundamental Markov property**. In colloquial terms, we say that the immediate future - represented by X_{n+1} - is independent of the entire past given the present.

For a Markov chain X_0, X_1, \dots the one-step transition probabilities are of paramount importance. These are the sequence of Markov kernels $(\hat{P}_{n,x})_{x \in \mathcal{X}}$, giving the conditional distributions of X_{n+1} given X_n . The fundamental Markov property shows that the kernel

$$(x_0, x_1, \dots, x_n) \mapsto \hat{P}_{n,x_n}$$

is in fact the conditional distribution of X_{n+1} given (X_0, X_1, \dots, X_n) . Hence we may write

$$\begin{aligned} P(X_0 \in A_0, X_1 \in A_1, \dots, X_n \in A_n) \\ = \int_{A_0 \times \dots \times A_{n-1}} \hat{P}_{n-1, x_{n-1}}(A_n) d(X_0, X_1, \dots, X_{n-1})(P)(x_0, x_1, \dots, x_{n-1}). \end{aligned}$$

But utilizing that $(\hat{P}_{n-2, x})_{x \in \mathcal{X}}$ by a slight change of the index set can be considered the conditional distribution of X_{n-1} given $(X_0, X_1, \dots, X_{n-2})$, we can by the extended Tonelli theorem write the above integral as a double integral:

$$\begin{aligned} P(X_0 \in A_0, X_1 \in A_1, \dots, X_n \in A_n) \\ = \int_{A_0 \times \dots \times A_{n-2}} \int_{A_{n-1}} \hat{P}_{n-1, x_{n-1}}(A_n) d\hat{P}_{n-2, x_{n-2}}(x_{n-1}) d(X_0, \dots, X_{n-2})(P)(x_0, \dots, x_{n-2}). \end{aligned}$$

And of course this process can be carried on, until we have the probability expressed as a n -fold integral:

$$\begin{aligned} P(X_0 \in A_0, X_1 \in A_1, \dots, X_n \in A_n) \\ = \int_{A_0} \int_{A_1} \dots \int_{A_{n-1}} \hat{P}_{n-1, x_{n-1}}(A_n) d\hat{P}_{n-2, x_{n-2}}(x_{n-1}) \dots d\hat{P}_{0, x_0}(x_1) d(X_0)(P)(x_0). \end{aligned}$$

In order to be slightly more specific, and avoid the indexing circus and the dots, an example of such a statement is

$$\begin{aligned} P(X_0 \in A, X_1 \in B, X_3 \in C, X_4 \in D) \\ = \int_A \int_B \int_C \hat{P}_{3, z}(D) d\hat{P}_{2, y}(z) d\hat{P}_{1, x}(y) dX_0(P)(x). \end{aligned}$$

So we have learned how to express the finite-dimensional distributions of a Markov chain through multiple integrals involving the one-step transition kernels. Believe it or not, this horrible characterization is usually taken as the definition of a Markov chain! For instance in Meyn and Tweedie (1993).

Sum up in some sense

It seems plausible to most people that this property generalizes certain facts about Markov Chains on a discrete space. But nobody has the slightest clue on how to check if it is satisfied. The literature abounds with statements that this or that collection of stochastic variables form a Markov chain, but there is never a proof - the

Markov property is taken as selfevident, even when it clearly is not. The problem is that noone will even know where to start, if they have to check that the finite-dimensional marginal distributions have an integral representation of the specified form. . . . It is way too complicated to be checkable in any practical sense. And hence the common conspiracy in the litterature: if everybody keeps quite, nobody will notice the problem. In Meyn and Tweedie (1993), where there are is a vast number of examples of Markov chains, not a single one of these examples is in fact proven to be Markovian! And this sad state of affairs is common in the litterature.

As we shall see, definition 2.1 can in fact be checked in a number of non-trivial situations, and so it represents a definite progress - we do not have to rely on divine insight when we claim processes to be Markovian.

Theorem 2.2 *If X_0, X_1, X_2, \dots is a Markov Chain, it holds that*

$$(X_n, X_{n+1}, \dots) \perp\!\!\!\perp (X_0, X_1, \dots, X_n) \mid X_n \quad \text{for all } n = 1, 2, \dots$$

PROOF: We show by induction on k that

$$(X_n, X_{n+1}, \dots, X_{n+k}) \perp\!\!\!\perp (X_0, X_1, \dots, X_n) \mid X_n \quad (2.2)$$

As the algebra

$$\bigcup_{k=1}^{\infty} \mathbb{F}(X_n, \dots, X_{n+k})$$

is a generator for $\mathbb{F}(X_n, X_{n+1}, \dots)$, stable under intersections, the extension of the result from the 'finite horizon future' to the 'infinite horizon future' follows from lemma 1.11.

To show (2.2) we observe that the statement for $k = 1$ is the very definition of the Markov chain (and for $k = 0$ it is downright triviality).

We know that

$$X_{n+k+1} \perp\!\!\!\perp (X_0, \dots, X_n, X_{n+1}, \dots, X_{n+k}) \mid X_{n+k}.$$

By shifting information to the righthand algebra to the conditioning algebra, we obtain that

$$X_{n+k+1} \perp\!\!\!\perp (X_0, \dots, X_n, X_{n+1}, \dots, X_{n+k}) \mid (X_n, \dots, X_{n+k}).$$

If we by induction assume that the property (2.2) is true for k , we have that combine via theorem 1.20 to obtain that

$$(X_n, X_{n+1}, \dots, X_{n+k}, X_{n+k+1}) \perp\!\!\!\perp (X_0, X_1, \dots, X_n) \mid X_n .$$

□

We usually refer to theorem 2.2 as the **general Markov property** - or simply as the Markov property. Colloquially speaking, the σ -algebra generated by (X_n, X_{n+1}, \dots) represents 'the future', and so the Markov property says that the future is independent of the past, given the present. What we have just proved is that if the immediate future only depends upon the past via the present at all times, then the general future will also depend upon the past via the present. Variations of the theme is clearly possible, for instance that

$$(X_0, X_1, \dots, X_m) \perp\!\!\!\perp (X_n, X_{n+1}, \dots) \mid (X_m, \dots, X_n) .$$

whenever $m < n$. This follows from shifting information around as we just did, followed by a reduction.

A formulation of the Markov property that is sometimes useful, and in fact by some authors is taken as the definition of a Markov Chain, is the following: if X_0, X_1, \dots is a Markov chain, and if $f : \mathcal{X}^{\mathbb{N}} \rightarrow \mathbb{R}$ is a bounded, measurable function, then for any n it holds that

$$E(f(X_n, X_{n+1}, \dots) \mid X_0, X_1, \dots, X_n) = E(f(X_n, X_{n+1}, \dots) \mid X_n) \quad \text{a.e}$$

This follows from combining theorem 2.2 and corollary 1.11. It is a nice property to have, and it is very flexible to work with. Used on functions like

$$(x_1, x_2, \dots) \mapsto 1_B(x_2)$$

it gives the fundamental Markovian property as a consequence. But considered as a definition, it has the same basic flaw as the definition via multiple integrals: nobody has a clue on how to check if it is satisfied in concrete examples.

Theorem 2.3 *Let Y_1, Y_2, \dots be independent variables, which we for simplicity assume have values in the same space $(\mathcal{Y}, \mathbb{K})$. Furthermore, let $\phi_n : \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{X}$ be a sequence of measurable maps.*

Let X_0 be yet another variable, independent of the Y 's, and define

$$X_n = \phi_n(X_{n-1}, Y_n) \quad \text{for } n = 1, 2, \dots \quad (2.3)$$

The proces X_0, X_1, \dots is a Markov chain.

PROOF: Due to independence, we have that

$$Y_{n+1} \perp\!\!\!\perp (X_0, Y_1, \dots, Y_n).$$

Which we might formulate as

$$Y_{n+1} \perp\!\!\!\perp (X_0, Y_1, \dots, Y_n) \mid \{\emptyset, \Omega\}.$$

As X_n is deterministically given by (X_0, Y_1, \dots, Y_n) , it is of course measurable with respect to the σ -algebra generated by these variables. And hence we may float it to the conditioning side,

$$Y_{n+1} \perp\!\!\!\perp (X_0, Y_1, \dots, Y_n) \mid X_n.$$

From there it may float back to the leftmost algebra, giving

$$(X_n, Y_{n+1}) \perp\!\!\!\perp (X_0, Y_1, \dots, Y_n) \mid X_n.$$

Now, X_{n+1} is (X_n, Y_{n+1}) -measurable, and X_0, X_1, \dots, X_n are all (X_0, Y_1, \dots, Y_n) -measurable. So by diminishing, we obtain that

$$X_{n+1} \perp\!\!\!\perp (X_0, X_1, \dots, X_{n-1}) \mid X_n$$

as desired. □

We usually refer to (2.3) as an **update scheme** for the Markov proces, and we refer to the Y -process as the **underlying error variables** or noise variables.

Theorem 2.4 *Let X_0, X_1, \dots be a Markov chain. There are update functions*

$$\phi_n : \mathcal{X} \times (0, 1) \rightarrow \mathcal{X} \quad \text{for all } n = 1, 2, \dots$$

with the following property: if U_1, U_2, \dots are a sequence of independent standard uniformly distributed stochastic variables, and if X'_0 is independent of the U 's, and has the same distribution as X_0 , then the update scheme

$$X'_n = \phi_n(X'_{n-1}, U_n) \quad \text{for } n = 1, 2, \dots$$

produces a proces X'_0, X'_1, X'_2, \dots with the same distribution as the original proces X_0, X_1, X_2, \dots

PROOF: We may represent the original chain by its initial distribution (the distribution of X_0) and each of its onestep transition kernels $(\hat{P}_{n,x})_{x \in \mathcal{X}}$. From these building blocks, we can build up the finite dimensional distributions of the process, and hence the joint distribution of all the entire process.

Each of the onestep transition kernels has an update function ϕ_n according to theorem 1.4. Using these in the update scheme above will produce a Markov chain with the same onestep transition kernels and then same initial distribution as the original chain, and hence the same overall distribution. □

So from a distributional point of view, we may always assume that a Markov chain is given by an update scheme - if a specific process, we happen to study, is not in update form, we can replace it by another process which is in update form, and which is indistinguishable from the first from a probabilistic point of view. The caveat is that the update functions are not in any way unique, and it may not be easy to produce update functions that make any sense intuitively.

The representations of Markov chains via update schemes is necessary for simulations purposes: a computer program that simulates a Markov chain must almost inevitably have form of an update scheme. But the idea also has a number of purely probabilistic applications.

One such application is to construct **couplings** of two Markov chains. A coupling is a realisation of the two processes on the same probability space. A trivial coupling is one where the two processes are independent. This can be achieved via update schemes, if the two processes have independent streams of error variables. But if we use the same stream of error variables, two heavily dependent Markov chains arise. To be specific, we may choose Y_0 and Z_0 in some way, independent of the U 's, and recursively define

$$Y_n = \phi_n(Y_{n-1}, U_n), \quad Z_n = \phi_n(Z_{n-1}, U_n) \quad \text{for } n = 1, 2, \dots$$

The two Markov chains will of course be heavily dependent, as they use the same error variables. But they coexist, and this coexistence may be utilized for various constructions - we may for instance observe that

$$(Y_n = Z_n) \subset (Y_{n+1} = Z_{n+1})$$

so if the processes at some stage are in the same point, they will move together hence forward. We may be interested in the probability that such a coalescence take place

sooner or later. A coupling with rather different properties can be obtained by

$$Y_n = \phi_n(Y_{n-1}, U_n), \quad Z_n = \phi_n(Z_{n-1}, 1 - U_n) \quad \text{for } n = 1, 2, \dots$$

In this construction, the chains will still be dependent, but they will not tend to coalesce.

Example 2.5 Time homogenous chains on finite state spaces. Given: typically the transition matrix. For simulation purposes: construct an update function.

◦

Example 2.6 The **random walk**, based on an iid. innovation sequence X_1, X_2, \dots , is by definition the stochastic process S_0, S_1, S_2, \dots given by

$$S_n = \sum_{i=1}^n X_i,$$

with the convention that $S_0 = 0$. This is a Markov chain with update scheme

$$S_n = S_{n-1} + X_n.$$

It is typically assumed that the innovations have mean zero, but random walks with positive (or negative) drift (meaning that the innovations have a nonzero mean) are study objects in their own right.

◦

Example 2.7 The **reflecting random walk**, based on an iid. innovation sequence X_1, X_2, \dots , is by definition the Markov chain with update scheme

$$T_0 = 0, \quad T_n = (T_{n-1} + X_n)^+.$$

A random walk with negative drift is frequently studied through the corresponding reflecting random walk, which exhibits the 'upwards excursions' of the random walk

◦

Example 2.8 The classical AR(1)-process on the real axis is given by the update scheme

$$X_{n+1} = \rho X_n + \epsilon_{n+1}$$

where the ϵ 's are independent and identically distributed. As a first choice, the errors are typically normally distributed with mean zero. But other choices are clearly

possible. We also have to specify the distribution of X_0 in order to specify the joint distribution of the proces.

In a sense, the behaviour of the AR(1)-proces is not very dependent on the specific choice of error distribution or initial distribution. The key is the magnitude of ρ . If $|\rho| < 1$, the proces will behave in a stable and quite predictable way. If $|\rho| > 1$ the proces will on the other hand explode. If $\rho = 1$ we are back in the random walk case. And if $\rho = -1$, we are essential also back in the random walk case, even though it becomes slighly more complicated to formulate the results. We willl return to this classification time and time again during the course.

◦

Example 2.9 There is a straight forward generalisation of the AR(1) proces to \mathbb{R}^k via the update scheme

$$X_n = RX_{n-1} + \epsilon_n$$

Here R is a $k \times k$ matrix, and the ϵ 's are an iid. sequence of \mathbb{R}^k -valued stochastic variables - a typical choice is to make the errors $\mathcal{N}(0, \Sigma)$ -distributed, where Σ is some legal variancematrix.

It is rather complicated to describe the long time behaviour of the chain, but at a first description it will depend on the eigenvalues of R . If all the eigenvalues are smallet than one in modulus, the matrix represents a linear map that contracts everything to 0. And this contraction is so dominating, that it even governs the stochastic behaviour. If some of the eigenvalues are outside of the complex unit circle, things become more complicate. The corresponding eigendirections will be 'directions of explosion', and they will in a sense govern the stochastic behaviour, unless the error distribution is so singular, that the proces will never have a non-zero component in an exploding direction.

Hence there is a very delicate interplay between the deterministic behavious of the underlying linear map, and the measuretheoretic singularities of the error distribution. At first sigth it would seem like a mathematical game to explore this interplay - it does not seem to be relevant from a modelling point of view. But it actually pops up in many places, and the problem must be considered seriously,

◦

Example 2.10 The AR(2)-process on the real axis is given by the update scheme

$$X_{n+1} = \alpha X_n + \beta X_{n-1} + \epsilon_{n+1}$$

where the ϵ 's are independent and identically distributed, typically normal with mean zero. As it stands, this update scheme does **not** give rise to a Markov proces, because it does not just depend on the present observation, but also a **lagged** observation. Furthermore, we need both X_0 and X_1 in order to be able to run the update mechanism.

But a slight rearrangement will in fact give a Markov chain. If we **stack** the proces, and consider the proces in \mathbb{R}^2 given by

$$Y_n = \begin{pmatrix} X_n \\ X_{n-1} \end{pmatrix},$$

we see that the Y -proces fits into the update scheme

$$\begin{aligned} Y_n &= \begin{pmatrix} \alpha X_{n-1} + \beta X_{n-2} + \epsilon_n \\ X_{n-1} \end{pmatrix} = \begin{pmatrix} \alpha & \beta \\ 1 & 0 \end{pmatrix} \begin{pmatrix} X_{n-1} \\ X_{n-2} \end{pmatrix} + \begin{pmatrix} \epsilon_n \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} \alpha & \beta \\ 1 & 0 \end{pmatrix} Y_{n-1} + \begin{pmatrix} \epsilon_n \\ 0 \end{pmatrix} \end{aligned}$$

This shows that the AR(2)-proces is just an AR(1)-proces in disguise, and hence it is 'practically Markovian'. The price we pay for this simplification is however, that the errors in the AR(1) updating scheme are quite degenerate - theyn are essentially one-dimensional. This perhaps sheds some light on the remarks as to why it is necessary to study AR(1)-processs in full generality, even with singular errordistributions.

o

Example 2.11 ARCH???? If yes, it must be here.

o

Example 2.12 Consider independent, identically distributed non-negative real random variables Y_1, Y_2, \dots , and think of them as representing **waiting times** between events. The occurrence of the n 'th event is thus happening at time

$$S_n = \sum_{i=1}^n Y_i.$$

The corresponding **renewal proces** is the continous time proces, which for each time-point indicates how many events that have occured,

$$N_t = \sup\{n \mid S_n \leq t\}$$

Renewal processes are very important in many branches of probability, in particular in Markov chain theory, and we will spend a considerable amount of time studying these objects.

But for now, we specialise to the situation where the waiting times have integer values. In that case we may introduce the **forward recurrence time chain**, V_1, V_2, \dots given by

$$V_n = \inf\{S_k - n \mid k \text{ such that } S_k > n\}$$

For any timepoint n , the value of V_n is the waiting time until the next event. If $V_n \geq 2$, there is no event taking place at time $n + 1$, and so $V_{n+1} = V_n - 1$. But if $V_n = 1$, there is an event taking place at time $n + 1$, and the value of V_{n+1} will be the length of the waiting period until the next event. Hence it is very easy to calculate the one-step probabilities:

$$P = \begin{pmatrix} \nu_1 & \nu_2 & \nu_3 & \dots \\ 1 & 0 & 0 & \dots \\ 0 & 1 & 0 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

where ν_1, ν_2, \dots are the pointmasses of waiting time distribution ν . But the relevance of the one-step probabilities are not clear, unless we know that the forward recurrence time chain is a Markov chain. And while that is true, a rigorous demonstration is not trivial. In example 2.27 we will establish Markovianess with blows and whistles, as a consequence of the so-called strong Markov property for the underlying random walk.

Backward recurrence time chains?

o

Example 2.13 Queues? If yes, it has to be here.

o

Some text describing the three different approaches to Markov chains: via kernels, via update schemes and the process-oriented approach we will adopt.

Some text on functions of Markov chains?

Example 2.14 If X_0, X_1, \dots is a Markov chain, and if $f : \mathcal{X} \rightarrow \mathbb{Y}$ is a measurable function, we may consider the process Y_0, Y_1, \dots given by

$$Y_n = f(X_n) \quad \text{for } n = 0, 1, 2, \dots$$

It is an important problem to find out if the Y -proces is Markovian as well. While reduction easily gives that

$$Y_{n+1} \perp\!\!\!\perp (Y_0, \dots, Y_{n-1}) \mid X_n,$$

there is no telling when we can shrink the conditioning algebra from $\mathbb{F}(X_n)$ to $\mathbb{F}(Y_n)$. The prominence of this problem arises, of course, from the fact that Y -proces is usually **not** markovian. It is actually rather difficult to find examples where the Markov property is preserved, but non-Markovianess is usually a mess to establish.

To construct an explicit example, we may let the X -proces be an asymmetric random walk on three points, say with one-step transition matrix

$$P = \begin{pmatrix} 0 & p & 1-p \\ 1-p & 0 & p \\ p & 1-p & 0 \end{pmatrix}$$

The initial distribution can be taken as the equidistribution. This process is a random movement on the corners of a triangle. When $p \neq \frac{1}{2}$, the process has a preoccupation for steps with a specific orientation. If p is close to one, the X -proces will move $1 \mapsto 2 \mapsto 3 \mapsto 1 \mapsto 2 \mapsto \dots$, if p is close to 0 the X -proces will move the other way around. As selftransitions are not possible, the variable (X_0, X_1, X_2) has only twelve non-zero pointmasses,

$$\begin{array}{ll} 1 & 2 & 1 & p(1-p)/3 & 2 & 3 & 1 & p^2/3 \\ 1 & 2 & 3 & p^2/3 & 2 & 3 & 2 & p(1-p)/3 \\ 1 & 3 & 1 & p(1-p)/3 & 3 & 1 & 2 & p^2/3 \\ 1 & 3 & 2 & (1-p)^2/3 & 3 & 1 & 3 & p(1-p)/3 \\ 2 & 1 & 2 & p(1-p)/3 & 3 & 2 & 1 & (1-p)^2/3 \\ 2 & 1 & 3 & (1-p)^2/3 & 3 & 2 & 3 & p(1-p)/3 \end{array}$$

The transformation we will consider is $f : \{1, 2, 3\} \rightarrow \{1, 2\}$ given by

$$f(1) = 1, f(2) = 2, f(3) = 2.$$

So the Y -proces is identical to the X -proces, except for the fact that the original states 2 and 3 are collapsed into one superstate, which for simplicity is called 2. The variable (Y_0, Y_1, Y_2) has five pointmasses (as 2-2 transitions are now perfectly legal, while 1-1 transitions are still forbidden),

$$\begin{array}{ll} 1 & 2 & 1 & 2p(1-p)/3 \\ 1 & 2 & 2 & p^2/3 + (1-p)^2/3 \\ 2 & 1 & 2 & p^2/3 + 2p(1-p)/3 + (1-p)^2/3 \\ 2 & 2 & 1 & p^2/3 + (1-p)^2/3 \\ 2 & 2 & 2 & 2p(1-p)/3 \end{array}$$

If we stratify this probability table by Y_1 , we get

$$\begin{array}{rcc}
 Y_1 = 1 & & \\
 & Y_2 = 1 & Y_2 = 2 \\
 Y_0 = 1 & 0 & 0 \\
 Y_0 = 2 & 0 & p^2/3 + 2p(1-p)/3 + (1-p)^2/3
 \end{array}$$

and

$$\begin{array}{rcc}
 Y_1 = 2 & & \\
 & Y_2 = 1 & Y_2 = 2 \\
 Y_0 = 1 & 2p(1-p)/3 & p^2/3 + (1-p)^2/3 \\
 Y_0 = 2 & p^2/3 + (1-p)^2/3 & 2p(1-p)/3
 \end{array}$$

There is actually independence in the $Y_1 = 1$ table, even if it is of a somewhat degenerate form. But there is no independence in the $Y_1 = 2$ table, unless $p = \frac{1}{2}$.

◦

2.2 The strong Markov property

The Markov property fomulates a relationship between the past, the present and the future, which is to hold for **all** values of 'the present', if a proces is to be called a Markov chain. At least it has to hold for all deterministic values. But it turns out time and time again, that we need the Markov property to hold in extended situationens, where the value of 'the present' is not known in advance, but has a certain stochasticity to it. As an example, we may consider 'the present' to be the first time, the proces enters a certain subset of \mathcal{X} .

To introduce the relevant formalism, we focus on a fixed proces X_0, X_1, \dots with values in some measurable space $(\mathcal{X}, \mathbb{E})$. The proces may or may not be a Markov chain, presently that is not relevant. The proces generates a **filtration**, a sequence of σ -algebras $\mathbb{F}_0 \subset \mathbb{F}_1 \subset \dots$ given by

$$\mathbb{F}_n = \mathbb{F}(X_0, X_1, \dots) \quad \text{for } n = 0, 1, 2, \dots$$

There is also a natural limit algebra \mathbb{F}_∞ , generated by all the variables X_0, X_1, \dots or - if we like - generated by the filtration. It may happen that \mathbb{F}_∞ equals the fundamental σ -algebra \mathbb{F} , but typically this is not the case - the fundamental algebra has to accommodate lots and lots of other stochastic variables, we draw upon at our convenience.

The process is of course **adapted** to the filtration, meaning that X_n is \mathbb{F}_n -measurable for each n .

A stochastic variable τ with values in the countable set $\mathbb{N}_0^* = \{0, 1, 2, \dots, \infty\}$ is called a **random time**. A random time is a **stopping time** with respect to the filtration $\mathbb{F}_0 \subset \mathbb{F}_1 \subset \dots$ if it satisfies that

$$(\tau = n) \in \mathbb{F}_n \quad \text{for } n = 0, 1, 2, \dots$$

The stopping time condition means that for each n there is a measurable subset $B_n \subset \mathcal{X}^{n+1}$ such that

$$(\tau = n) = \left((X_0, X_1, \dots, X_n) \in B_n \right).$$

The implication is that we are able to read off from the values of X_0, X_1, \dots, X_n whether $\tau = n$ or not. By observing the X -process for some time, we know if τ has occurred or not.

Example 2.15 Deterministic stopping times.

First hitting times.

Combinations: minima, maxima.

First hitting time **after** a given stopping time.

First entrance time, first return time.

o

It is in principle allowed that a stopping time τ can obtain the value ∞ . In martingale theory this is not only a sensible convention, but in fact a useful idea, that vastly simplifies a number of formulations. But in Markov chain theory, infinite stopping times are a menace, and we will usually not allow them. We will focus on three types of stopping times: The **bounded** stopping times, which never take on values above a certain threshold known to us, The **finite** stopping times, which never take on the value ∞ , but may take on arbitrarily large integral values. And the **almost surely finite** stopping times, which satisfy that

$$P(\tau < \infty) = 1.$$

We would really like all our stopping times to be finite - but that would exclude the first hitting times from considerations. Consider the waiting times until head

comes up in a coin tossing experiment. With probability one, head comes up sooner or later. But there is a formal possibility that head never comes up, and we have to deal with this possibility in our formalism. We could cut the corresponding nullset out of the background probability space Ω , to ensure that head always comes up. But if we follow that route, we will have to do this kind of surgery on the background space whenever we introduce a new stopping time, and it becomes technically very unpleasant in the long run. It is much neater to allow that stopping times take on the value ∞ - as long as we sure this only happens on a nullset.

If X_0, X_1, \dots is a proces and τ is a corresponding stopping time, we introduce the symbol X_τ as the value of the proces at the random time τ . If τ is finite, the formal definition may be written as

$$X_\tau = \sum_{n=0}^{\infty} 1_{(\tau=n)} X_n$$

but to a certain extend, this breaks down if the stopping time can take on the value ∞ - even if this only happens with probability zero. In order to do something, we adopt the convention that whenever we introduce a new measurable space $(\mathcal{X}, \mathbb{E})$ on which stochastic variables may have values, we equip it with a standard variable X^* - on \mathbb{R}^n we could let the standard variable have the deterministic value 0. We will assume that this standard variable is measurable with respect to \mathbb{F}_∞ but no other details matter. Having introduced a standard variable, we may then define

$$X_\tau = \sum_{n=0}^{\infty} 1_{(\tau=n)} X_n + 1_{(\tau=\infty)} X^* .$$

If τ assumes the value ∞ with positive probability, the choice of standard variable is of course important for the behaviour of X_τ . But as long as τ is almost surely finite, the invention of the standard variable is a purely formal gimmick. Observe that X_τ becomes measurable with respect to \mathbb{F}_∞ :

$$\begin{aligned} (X_\tau \in A) &= \bigcup_{n=0}^{\infty} (X_n \in A) \cap (\tau = n) \quad \cup \quad (X^* \in A) \cap (\tau = \infty) \\ &= \bigcup_{n=0}^{\infty} (X_n \in A) \cap (\tau = n) \quad \cup \quad (X^* \in A) \cap (\tau = \infty) \end{aligned}$$

The only event in this compositon that does not obviously satisfy the relevant measurability condition is $(\tau = \infty)$. But the complement $(\tau < \infty)$ is the union of event of the form $(\tau = n)$, and this establishes measurability.

Corresponding to a stopping time τ , we have a natural notion of 'the past', namely the σ -algebra

$$\mathbb{F}_\tau = \{F \in \mathbb{F} \mid F \cap (\tau = n) \in \mathbb{F}_n \text{ for all } n = 0, 1, \dots\}.$$

Lemma 2.16 *Let X_0, X_1, \dots be a stochastic process, and let τ be an adapted stopping time. Then the variables τ and X_τ are both \mathbb{F}_τ -measurable.*

PROOF: Trivial manipulations. If we consider the event $(\tau = k)$, we have that

$$(\tau = k) \cap (\tau = n) = \begin{cases} (\tau = n) & \text{if } k = n \\ \emptyset & \text{if } k \neq n \end{cases}$$

In both cases we get that $(\tau = k) \cap (\tau = n) \in \mathbb{F}_n$. And hence $(\tau = k) \in \mathbb{F}_\tau$. This shows the measurability of τ .

Similarly, if we let A be a measurable subset of \mathcal{X} , we have that

$$(X_\tau \in A) \cap (\tau = n) = (X_n \in A) \cap (\tau = n) \in \mathbb{F}_n,$$

so $(X_\tau \in A) \in \mathbb{F}_\tau$. □

Lemma 2.17 *Let X_0, X_1, \dots be a stochastic process, and let τ and σ be two adapted stopping times. It holds that*

$$\sigma \leq \tau \quad \Rightarrow \quad \mathbb{F}_\sigma \subset \mathbb{F}_\tau.$$

PROOF: This is well known. □

Lemma 2.18 *Let X_0, X_1, \dots be a Markov chain, with corresponding filtration $\mathbb{F}_0 \subset \mathbb{F}_1 \subset \dots$. If Z and W are two bounded real variables, both \mathbb{F}_n -measurable, then it holds that*

$$E(Z \mid X_n) = E(W \mid X_n) \text{ a.e.} \quad \Rightarrow \quad E(Z \mid X_n, X_{n+1}) = E(W \mid X_n, X_{n+1}) \text{ a.e.}$$

PROOF: This is really a trivial consequence of the Markov property. The future variable X_{n+1} is independent of the past algebra \mathbb{F}_n , in particular of Z and W , given the present variable X_n . Referring to the asymmetric formulation of conditional independence in corollary 1.11, we obtain the string of equations

$$E(Z | X_n, X_{n+1}) = E(Z | X_n) = E(W | X_n) = E(W | X_n, X_{n+1}) \quad \text{a.e.}$$

□

Note the amusing fact that we are somehow using the Markov property backwards in this proof. The argument can be verbalized as saying that when we are attempting to 'predict the past', there is no information in knowing the future - only the present matters.

Lemma 2.19 *Let X_0, X_1, \dots be a proces, with corresponding filtration $\mathbb{F}_0 \subset \mathbb{F}_1 \subset \dots$, and let τ be an adapted stopping time. Let Z be a realvalued stochastic variable, measurable with respect to \mathbb{F}_τ . For any $n < \infty$ it holds that $1_{(\tau=n)}Z$ is m measurable with respect to \mathbb{F}_n .*

PROOF: We simply observe that for any $B \in \mathbb{B}$ we have one of two situations, depending on whether B contains 0 or not:

$$\left(1_{(\tau=n)}Z \in B\right) = \begin{cases} (Z \in B) \cap (\tau = n) & 0 \notin B \\ (Z \in B) \cap (\tau = n) \cup (\tau \neq n) & 0 \in B \end{cases}$$

Since Z is assumed to be \mathbb{F}_τ -measurable, $(Z \in B)$ will be an \mathbb{F}_τ -set, and so $(Z \in B) \cap (\tau = n)$ will be an \mathbb{F}_n -set. Also the event $(\tau \neq n)$ is \mathbb{F}_n -measurable - its complement has the relevant measurability per definition. So in either case $\left(1_{(\tau=n)}Z \in B\right)$ is in \mathbb{F}_n . □

Lemma 2.20 *Let X_0, X_1, \dots be a proces, with corresponding filtration $\mathbb{F}_0 \subset \mathbb{F}_1 \subset \dots$, and let τ be an adapted stopping time. For any event F and any $n < \infty$ it holds that*

$$E\left(1_{(\tau=n)}P(F | X_\tau, \tau) | X_n\right) = P(F \cap (\tau = n) | X_n) \quad \text{a.e} \quad (2.4)$$

PROOF: The claim that two conditional expectations with respect to X_n are the same, of course means that the two stochastic variables integrate to the same thing, when integrated over $\mathbb{F}(X_n)$ -events. Observe that

$$\begin{aligned} \int_{(X_n \in A)} 1_{(\tau=n)} P(F | X_\tau, \tau) dP &= \int_{(X_\tau \in A, \tau=n)} P(F | X_\tau, \tau) dP \\ &= P(F \cap (X_\tau \in A, \tau = n)), \end{aligned}$$

since the middle integral is over an $\mathbb{F}(X_\tau, \tau)$ -event. Similarly it holds that

$$\int_{(X_n \in A)} 1_{F \cap (\tau=n)} dP = P(F \cap (X_n \in A) \cap (\tau = n)) = P(F \cap (X_\tau \in A, \tau = n)).$$

□

Note that (2.4) may be formulated

$$E(P(F \cap (\tau = n) | X_\tau, \tau) | X_n) = P(F \cap (\tau = n) | X_n) \quad \text{a.e.}$$

since $1_{(\tau=n)}$ is $\mathbb{F}(X_\tau, \tau)$ -measurable. Hence we see that the statement is really about a double conditioning situation. The statement remains non-trivial, however, because the σ -algebras in question, \mathbb{F}_n and $\mathbb{F}(X_\tau, \tau)$, are not nested. In fact, the statement is only true due to the specific nature of the event $F \cap (\tau = n)$ and its interplay with the two σ -algebras.

Theorem 2.21 (Strong Markov property) *Let X_0, X_1, \dots be a Markov chain, with corresponding filtration $\mathbb{F}_0 \subset \mathbb{F}_1 \subset \dots$. Let τ be an adapted stopping time, and assume that τ is almost surely finite. Then*

$$X_{\tau+1} \perp\!\!\!\perp \mathbb{F}_\tau | (\tau, X_\tau)$$

PROOF: We prove that for any $F \in \mathbb{F}_\tau$ it holds that

$$P(F | X_\tau, X_{\tau+1}, \tau) = P(F | X_\tau, \tau) \quad \text{a.e.} \quad (2.5)$$

which is another instance of the the-future-is-irrelevant-for-predicting-the-past phenomenon, we have previously encountered. The righthand side of (2.5) clearly has

the measurability properties to be a version of the lefthand side, so we only need to check that it has the right integrals over $\mathbb{F}(X_\tau, X_{\tau+1}, \tau)$ -events. For finite n we see that

$$\int_{(\tau=n, X_\tau \in A, X_{\tau+1} \in B)} P(F | X_\tau, \tau) dP = \int_{(X_n \in A, X_{n+1} \in B)} 1_{(\tau=n)} P(F | X_\tau, \tau) dP$$

Combining the lemmas, we see that we can replace the integrand by $1_{(\tau=n) \cap F}$ to obtain

$$\begin{aligned} \int_{(\tau=n, X_\tau \in A, X_{\tau+1} \in B)} P(F | X_\tau, \tau) dP &= P((X_n \in A, X_{n+1} \in B, \tau = n) \cap F) \\ &= P((X_\tau \in A, X_{\tau+1} \in B) \cap (\tau = n) \cap F) \end{aligned}$$

It is trivially true that

$$\int_{(\tau=\infty, X_\tau \in A, X_{\tau+1} \in B)} P(F | X_\tau, \tau) dP = P((X_\tau \in A, X_{\tau+1} \in B) \cap (\tau = \infty) \cap F)$$

since both sides are zero, due to the assumption that τ is almost surely finite. The events of the form $(\tau = n, X_\tau \in A, X_{\tau+1} \in B)$ (including the events with $n = \infty$) form a generator for $\mathbb{F}(X_\tau, X_{\tau+1}, \tau)$, stable under the formation of intersections. And hence it follows that

$$\int_G P(F | X_\tau, \tau) dP = P(G \cap F) \quad \text{for all } G \in \mathbb{F}(X_\tau, X_{\tau+1}, \tau).$$

That is, we have established (2.5). □

This 'strong Markov property' is slightly weaker than we would have liked. The immediate future only becomes independent of the past given the present if 'the present' includes a glance at the clock. One can construct examples which shows that in general the information of the random time cannot be dispensed of. But in the next section we will go hunting for a setup, where all times look the same, and where there is no essential information in knowing the value of τ . In that setup we will be able to strengthen the conclusion in theorem 2.21 to obtain what is generally perceived as the strong Markov property in the litterature.

Example 2.22 Asymmetric random walk on three points with a direction which oscillates back and forth. One-step transition matrices are

$$P_{2n-1} = \begin{pmatrix} 0 & p & 1-p \\ 1-p & 0 & p \\ p & 1-p & 0 \end{pmatrix}, \quad P_{2n} = \begin{pmatrix} 0 & 1-p & p \\ p & 0 & 1-p \\ 1-p & p & 0 \end{pmatrix}$$

Starting distribution: equidistribution on state 2 and 3. Stopping time: first hitting time of state 1. In that case $X_\tau = 1$, and so no information is contained in that variable. The strong Markov property asserts that $X_{\tau+1}$ is independent of X_1 given τ . Is it possible to show (though perhaps not easily) that $X_{\tau+1}$ is **not** independent of X_1 unconditionally.

◦

Corollary 2.23 *Let X_0, X_1, \dots be a Markov chain, with corresponding filtration $\mathbb{F}_0 \subset \mathbb{F}_1 \subset \dots$. Let τ be an adapted stopping time, and assume that τ is almost surely finite. Then*

$$(X_{\tau+1}, X_{\tau+2}, \dots) \perp\!\!\!\perp \mathbb{F}_\tau \mid (\tau, X_\tau) \quad (2.6)$$

PROOF: We show by an induction argument that

$$(X_{\tau+1}, X_{\tau+2}, \dots, X_{\tau+k}) \perp\!\!\!\perp \mathbb{F}_\tau \mid (\tau, X_\tau) \quad (2.7)$$

for all values of k . The crux of the matter is that $\sigma = \tau + k$ is a stopping time. Hence

$$X_{\tau+k+1} \perp\!\!\!\perp \mathbb{F}_{\tau+k} \mid (\tau + k, X_{\tau+k})$$

As the variables $X_\tau, X_{\tau+1}, \dots, X_{\tau+k}$ are all $\mathbb{F}_{\tau+k}$ -measurable, we can shift them to the conditioning algebra, and obtain

$$X_{\tau+k+1} \perp\!\!\!\perp \mathbb{F}_{\tau+k} \mid (\tau + k, X_\tau, X_{\tau+1}, \dots, X_{\tau+k})$$

As $\tau + k \geq \tau$ we see that $\mathbb{F}_\tau \subset \mathbb{F}_{\tau+k}$, so by shrinking it follows that

$$X_{\tau+k+1} \perp\!\!\!\perp \mathbb{F}_\tau \mid (\tau + k, X_\tau, X_{\tau+1}, \dots, X_{\tau+k})$$

Observing that τ and $\tau + k$ generate the same σ -algebra (note: the stopping time algebras are not the same, but here we are talking about the algebras that make the variables measurable), and combining with the inductive hypothesis (2.7) we obtain via theorem 1.20 that

$$(X_{\tau+1}, X_{\tau+2}, \dots, X_{\tau+k+1}) \perp\!\!\!\perp \mathbb{F}_\tau \mid (\tau, X_\tau)$$

as desired.

□

2.3 Homogeneity

Virtually every single Markov chain we will consider, will have a further simplifying property called **time homogeneity**.

Homogeneity in terms of transition probabilities: The usual definition is that we can pick the one-step transition kernels so that they are all the same. There is one Markov kernel that satisfies

$$P(X_n \in A, X_{n+1} \in B) = \int_A \hat{P}_x(B) dX_n(P)(x) \quad \text{for all } A, B \text{ and } n.$$

Note the inherent difficulty in this definition: there are many ways to pick the various 1-step transition kernels, and if these are not picked *in concerto* they will surely differ.

The obvious example exhibiting the problems is the random walk, with symmetric ± 1 increments. It is a time homogenous Markov chain with transition matrix

$$p_{nm} = \begin{cases} \frac{1}{2} & \text{if } m = n + 1 \\ \frac{1}{2} & \text{if } m = n - 1 \\ 0 & \text{otherwise} \end{cases}$$

This is the obvious transition matrix that everybody will write down - before they start thinking. But there are lots of other choices. Usually we insist that the random walk starts in 0. If that is the case, the transition matrix for the time 2 to time 3 transition is only uniquely given from the states $-2, 0$ and 2 . Similarly, the transition matrix for the time 3 to time 4 transition is only uniquely given from the states $-3, -1, 1$ and 3 . Transitions from all other states are not determined at all. So if we pick the transition matrices one by one, it is quite unlikely that we will pick the same every time, unless we have some principle to guide us.

Homogeneity in terms of update schemes: If Y_1, Y_2, \dots are independent **and identically distributed** and if we utilize the same update function $\phi : \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{X}$ in every step, then the chain generated as

$$X_{n+1} = \phi(X_n, Y_{n+1})$$

will be time homogenous in the sense that the obvious one-step transition kernels, formed from the distribution of the Y 's and from ϕ by the substitution theorem, are

all the same. In the other direction, it is also clear that if a Markov chain where all the one-step transition kernels are the same, then there is an update scheme of the above sort generating the process.

The examples we gave of Markov chains with specified update schemes were all of this time homogeneous form. Actually, time-varying update schemes virtually never appear in applications. With one notable exception: **simulated annealing** which is an optimization algorithm based on Markov chains.

However, for processes constructed on top of other processes, neither the Markov structure nor the time homogeneity may be immediately visible. We have shown that we may examine the Markov property from first principles - but the random walk example above shows that we have to be very careful when we check for time homogeneity. We adopt the following slightly weaker definition:

Definition 2.24 A Markov chain X_0, X_1, \dots is **weakly time homogeneous** if

$$X_{\tau+1} \perp\!\!\!\perp \tau \mid X_\tau$$

for every adapted, almost surely finite stopping time τ .

This definition undoubtedly looks confusing. There is a nice linguistic catch in that homogeneity with this definition reflects that something is 'independent of time' in a stochastic sense. But apart from that, the definition may seem arbitrary. However, the definition has its merits:

The concept of weak time homogeneity is a distributional concept

It is enough to check weak time homogeneity on finite stopping times. If not on bounded stopping times.

Theorem 2.25 A time homogeneous Markov chain X_0, X_1, \dots is weakly time homogeneous.

PROOF: We may assume that the Markov chain has update form,

$$X_{n+1} = \phi(X_n, U_{n+1})$$

for some fixed map ϕ , and a sequence of independent, standard uniformly distributed real stochastic variables U_1, U_2, \dots . Let τ be a finite stopping time. Then

$$X_{\tau+1} = \sum_{n=0}^{\infty} 1_{(\tau=n)} X_{n+1} = \sum_{n=0}^{\infty} 1_{(\tau=n)} \phi(X_n, U_{n+1}) = \phi(X_\tau, \tilde{U}) \quad (2.8)$$

where we have introduced the variable

$$\tilde{U} = \sum_{n=0}^{\infty} 1_{(\tau=n)} U_{n+1}.$$

Take an event $F \in \mathbb{F}_{tau}$. Then it holds that

$$P((\tilde{U} \in A) \cap F) = \sum_{n=0}^{\infty} P((\tilde{U} \in A) \cap F \cap (\tau = n)) = \sum_{n=0}^{\infty} P((U_{n+1} \in A) \cap F \cap (\tau = n)).$$

Using that $F \cap (\tau = n)$ is \mathbb{F}_n -measurable, and that U_{n+1} is independent of \mathbb{F}_n , as this algebra is contained in $\mathbb{F}(X_0, U_1, \dots, U_n)$, we get that

$$\begin{aligned} P((\tilde{U} \in A) \cap F) &= \sum_{n=0}^{\infty} P(U_{n+1} \in A) P(F \cap (\tau = n)) \\ &= P(U_1 \in A) \sum_{n=0}^{\infty} P(F \cap (\tau = n)) \\ &= P(U_1 \in A) P(F). \end{aligned}$$

We can draw two consequences: For one thing, \tilde{U} is standard uniformly distributed. But more important: we see that

$$\tilde{U} \perp \mathbb{F}_\tau.$$

Observing that X_τ is \mathbb{F}_τ -measurable, we may float information to the (trivial) conditioning side, and obtain that

$$\tilde{U} \perp \mathbb{F}_\tau | X_\tau.$$

We may float information back, and obtain

$$(\tilde{U}, X_\tau) \perp \mathbb{F}_\tau | X_\tau.$$

As $X_{\tau+1}$ according to (2.8) is $\mathbb{F}(X_\tau, \tilde{U})$ -measurable, and τ is \mathbb{F}_τ -measurable, it follows by reduction that

$$X_{\tau+1} \perp \tau | X_\tau,$$

as desired.

□

A nice consequence of the proof is that the conditional distribution of $X_{\tau+1}$ given X_τ is simply the same as the common conditional distribution of X_{n+1} given X_n , since it follows the same update rule with an error variable that is standard uniformly distributed.

Theorem 2.26 (Strong Markov property) *Let X_0, X_1, \dots be a Markov chain, with corresponding filtration $\mathbb{F}_0 \subset \mathbb{F}_1 \subset \dots$. Assume that the chain is weakly time homogeneous. Let τ be an adapted stopping time, and assume that τ is almost surely finite. Then*

$$X_{\tau+1} \perp\!\!\!\perp \mathbb{F}_\tau \mid X_\tau$$

PROOF: We combine weak homogeneity and theorem 2.21 via theorem 1.20, the result drops out for free.

□

This version of the strong Markov property is perhaps the key property of time homogeneous Markov chains in any formulation.

Example 2.27 Apply the strong Markov property to show that the forward recurrence time chain for a renewal process is in fact a Markov chain. *****

○

2.4 The Chapman-Kolmogorov equations

Definition of composition of kernels

Associativity

Powers of a Markov kernel

The full transition structure

The Chapman-Kolmogorov equations

Interpretations

2.5 Sampled chains

The inhomogenous case.

The homogenous case.

The counterexample.