Markov Chains and Mixing Times

Jan M. Swart (Czech Academy of Sciences)

Thursday, October 1st, 2020

The Markov property

Lemma For random variables $X_0, \ldots, X_n$ taking values in a finite set $\mathcal{X}$, the following conditions are equivalent:

(i) For each $0 < t < n$, the random variables $(X_0, \ldots, X_{t-1})$ and $(X_{t+1}, \ldots, X_n)$ are conditionally independent given X_t.

(ii) For each $0 < t \leq n$, there exists a probability kernel $P_{t-1,t}$ such that $\mathbb{P}[X_t = x \mid X_0, \ldots, X_{t-1}] = P_{t-1,t}(X_{t-1}, x)$ a.s. for all $x \in \mathcal{X}$.

(iii) There exists a probability law μ on $\mathcal{X}$ and probability kernels $(P_{t-1,t})_{0<t\leq n}$ such that $\mathbb{P}[(X_0, \ldots, X_n) = (x_0, \ldots, x_n)] = \mu(x_0)P_{0,1}(x_0, x_1) \cdots P_{n-1,n}(x_{n-1}, x)$ for all $x_0, \ldots, x_n \in \mathcal{X}$.

We say that $(X_t)_{0\leq t\leq n}$ is a *Markov chain* with *initial law* μ and *transition kernels* $(P_{t-1,t})_{0<t\leq n}$. If it is possible to choose the transition kernels such that $P(x, y) = P_{t-1,t}(x, y)$ does not depend on t, then the Markov chain is *time-homogenous*.

Markov chains

Usually, the starting point is *not* a sequence of random variables $X_0, \ldots, X_n$, but a transition kernel P. We fix P and are interested in Markov chains with this transition kernel (and arbitrary initial law). We write $\mathbb{P}_\mu$ (resp. $\mathbb{P}_x$) for the law of the Markov chain with initial law μ (resp. δ_x).

Using Kolmogorov's extension theorem, we can without loss of generality take $n = \infty$, so usually we consider Markov chains $(X_t)_{t\geq 0}$.

One can idealize further and allow $\mathcal{X}$ to be countably infinite.
With more work, one can allow uncountable $\mathcal{X}$.
It is also possible to consider continuous time.
This leads to the general theory of *Markov processes*, which describe limits of finite Markov chains with large state spaces.

In the book, we will stick to *finite* state space $\mathcal{X}$, but we will nevertheless be interested in *large* $\mathcal{X}$.

Matrix notation

We observe that

$$\begin{aligned}\mathbb{P}_x[X_2 = z] &= \sum_y \mathbb{P}_x[X_1 = y,\ X_2 = z] \\ &= \sum_y P(x,y)P(y,z) = P^2(x,z).\end{aligned}$$

More generally

$$\mathbb{P}_\mu[X_n = y] = \sum_x \mu(x)P^n(x,y) =: \mu P^n(y).$$

Also

$$\mathbb{E}_x[f(X_n)] = \sum_y P^n(x,y)f(y) =: P^n f(x).$$

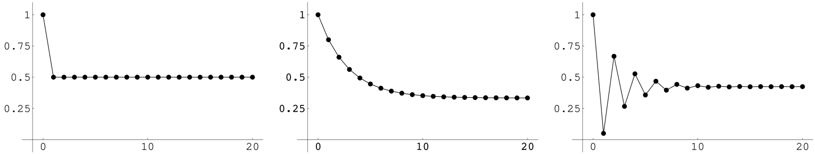
We view probability laws as *row vectors*, transition kernels as *matrices*, and real functions as *column vectors*.

A simple example

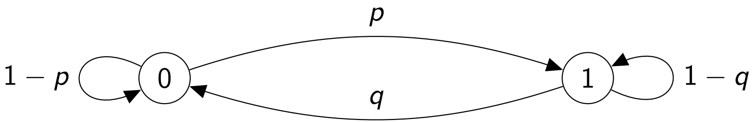
Consider

$$\mathbb{P} := \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}.$$

We plot $P^t(1,1)$ as a function of t:



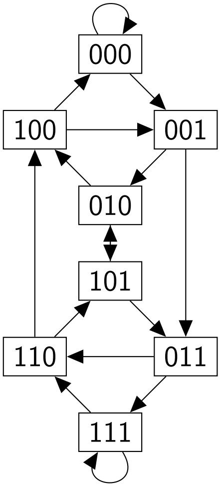
$p = q = 1/2$ $\quad$ $p = 0.2,\ q = 0.1$ $\quad$ $p = 0.95,\ q = 0.7$



Another example

Let $(Z_t)_{t\geq 0}$ be i.i.d. Bernoulli random variables with $\mathbb{P}[Z_t = 0] = \mathbb{P}[Z_t = 1] = \frac{1}{2}$. Then $(Z_t, Z_{t+1}, Z_{t+2})_{t\geq 0}$ is a Markov chain with state space $\mathcal{X} = \{0,1\}^3$.

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$



Random mapping representations

Lemma
Let $(Z_t)_{t\geq 1}$ be i.i.d. random variables with values in Λ.
Let X_0 be an independent random variable with values in $\mathcal{X}$.
Let $f : \mathcal{X} \times \Lambda \to \mathcal{X}$ be a (measurable) function. Then

$$X_t := f(X_{t-1}, Z_t) \qquad (t \geq 0)$$

defines a Markov chain with transition kernel

$$P(x, y) = \mathbb{P}[f(x, Z_1) = y] \qquad (x, y \in \mathcal{X}).$$

Proof By induction, Z_t is independent of $X_0, \ldots, X_{t-1}$, hence

$$\begin{aligned}
&\mathbb{P}[X_t = x_t \,|\, (X_0, \ldots, X_{t-1}) = (x_0, \ldots, x_{t-1})] \\
&\quad = \mathbb{P}[f(x_{t-1}, Z_t) = x_t \,|\, (X_0, \ldots, X_{t-1}) = (x_0, \ldots, x_{t-1})] \\
&\quad = \mathbb{P}[f(x_{t-1}, Z_t) = x_t] = P(x_{t-1}, x_t).
\end{aligned}$$

Random mapping representations

Setting $M(x) := f(x, Z)$ defines a random map $M : \mathcal{X} \to \mathcal{X}$ such that

$$(RM) \qquad\qquad P(x, y) = \mathbb{P}[M(x) = y].$$

Conversely, if M is defined on a probability space Ω, then $f(x, \omega) := M(\omega)(x)$ is a map $f : \mathcal{X} \times \Omega \to \mathcal{X}$.

Every probability kernel has a random mapping representation:
For each $x \in \mathcal{X}$, let $M(x)$ be a random variable with law $P(x, \cdot\,)$.
Couple the random variables $(M(x))_{x \in \mathcal{X}}$ in any way.
Then (RM) holds.

Random mapping representations are not unique, since we are free to choose any joint law for $(M(x))_{x \in \mathcal{X}}$,
as long as the marginals satisfy (RM).

Random mapping representations

Simulations of Markov chains on a computer usually involve a random mapping representation with $(Z_t)_{t\geq 1}$ i.i.d. uniformly distributed on $[0,1]$.

Example $\mathcal{X} = \{0,1\}$, $P(x,y) = P(y,x) = \frac{1}{2}$.

Representation 1 $f(x,Z) := 1-x$ if $Z \leq \frac{1}{2}$ and $f(x,Z) := x$ otherwise.

Representation 2 $f(x,Z) := 1$ if $Z \leq \frac{1}{2}$ and $f(x,Z) := 0$ otherwise.

Random mapping representations yield a natural way of coupling Markov chains with different initial states.
Let $(X^x_t)_{t\geq 0}$ be the Markov chain with initial state $X^x_0 = x$.

In Representation 1, $\mathbb{P}[X^0_t \neq X^1_t] = 1$ for all $t \geq 1$.
In Representation 2, $\mathbb{P}[X^0_t \neq X^1_t] = 0$ for all $t \geq 1$.

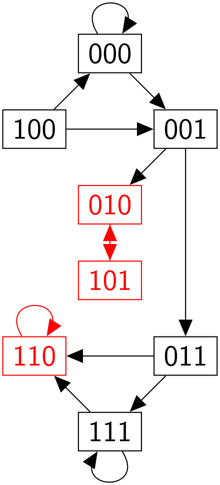
Classification of states

Write $x \longrightarrow y$ if $P^t(x,y) > 0$ for some $t \geq 0$.
In the oriented graph picture, this means that there is a walk from x to y.

Write $x \longleftrightarrow y$ if $x \longrightarrow y \longrightarrow x$. This defines an equivalence relation. The equivalence classes are called *communicating classes*.

A communicating class C is *essential* if there are no $x \in C$, $y \notin C$ such that $x \to y$.
An essential class with one element is an absorbing state.

A Markov chain is *irreducible* if $\mathcal{X}$ is a communicating class.



Aperiodicity

By definition, the *period* of x is the greatest common divisor of $\{t \geq 1 : P^t(x,x) > 0\}$.

- All states in a communicating class have the same period.
- An irreducible chain with period 1 is called *aperiodic*.
- If an irreducible chain has period n, then we can partition $\mathcal{X} = \mathcal{X}_0 \cup \cdots \cup \mathcal{X}_{n-1}$ so that from $\mathcal{X}_k$, it is only possible to jump to $\mathcal{X}_{k+1 \bmod (n)}$.
- If $(X_t)_{t\geq 0}$ has period n, then $(X_{nt})_{t\geq 0}$ is aperiodic with state space $\mathcal{X}_0$.
- If $\mathcal{X}$ is finite and P is irreducible and aperiodic, then there exists a $t > 0$ such that $P^t(x,y) > 0$ for all $x, y \in \mathcal{X}$.

Hitting times

For $x \in \mathcal{X}$, we define:

$$\text{hitting time} \qquad \tau_x := \inf\{t \geq 0 : X_t = x\},$$
$$\text{first return time} \quad \tau_x^+ := \inf\{t \geq 1 : X_t = x\}.$$

Lemma In an irreducible chain with finite state space,

$$\mathbb{E}_x[\tau_y] < \infty \quad \text{and} \quad \mathbb{E}_x[\tau_y^+] < \infty \qquad (x, y \in \mathcal{X}).$$

Proof In the aperiodic case, we can choose $t > 0$ such that $P^t(x, y) > 0$ for all $x, y \in \mathcal{X}$.
Fix y and set $\varepsilon := \inf_{x \in \mathcal{X}} P^t(x, y)$. Then

$$\mathbb{P}[X_{nt} \neq y \,|\, X_t \neq y,\ X_{2t} \neq y, \ldots, X_{(n-1)t} \neq y] \leq 1 - \varepsilon,$$

and hence

$$\mathbb{P}_x[\tau_y^+ > nt] \leq (1 - \varepsilon)^n,$$

so

$$\mathbb{E}_x[\tau_y] \leq \mathbb{E}_x[\tau_y^+] = \sum_{t \geq 0} \mathbb{P}_x[\tau_y^+ > t] < \infty. \qquad \blacksquare$$

Transient states

Lemma If C is an inessential class and

$$\sigma_C := \inf\{t \geq 0 : X_t \notin C\},$$

is the *first exit time* of C, then $\mathbb{E}_x[\sigma_C] < \infty$ for all $x \in C$.

Proof Define a new Markov chain with state space $\mathcal{Y} := C \cup \{*\}$ and transition kernel $Q(x, y) := P(x, y)$ $(x, y \in C)$, $Q(x, *) := \sum_{y \in \mathcal{X} \setminus C} P(x, y)$, and $Q(*, x) := 1$ for some fixed $x \in C$. This new chain is irreducible and σ_C is equally distributed with τ_*. ∎

Invariant laws

An *invariant law* is a probability law π that satisfies the equivalent conditions:

- $\pi P = \pi$,
- $\mathbb{P}_\pi$ is *stationary*, i.e.,
  $\mathbb{P}_\pi[(X_1, \ldots, X_t) \in \cdot\,] = \mathbb{P}_\pi[(X_0, \ldots, X_{t-1}) \in \cdot\,] \quad \forall t.$

Proposition On each essential class C, there exists a unique invariant law. On inessential classes, there do not exist invariant laws.

Proof idea Fix a reference point $x \in C$. By definition, an *excursion away from* x is a finite sequence $\vec{x} = (x_0, \ldots, x_n)$ with $n \geq 1$, $x_0 = x = x_n$, and $x_k \neq x$ for all $0 < k < n$. Define

$$\mu(x_0, \ldots, x_n) := \mathbb{P}_x\big[(X_0, \ldots, X_{\tau_x^+}) = (x_0, \ldots, x_n)\big].$$

Invariant laws

Assume that $(X_t)_{t\in\mathbb{Z}}$ is stationary. Let

$$\sigma_x := \inf\{t \leq -1 : X_t = x\} \quad \text{and} \quad \tau_x := \inf\{t \geq 0 : X_t = x\}.$$

Then

$$\begin{aligned}&\mathbb{P}\big[\sigma_x = -t,\ (X_{\sigma_x}, \ldots, X_{\tau_x}) = (x_0, \ldots, x_n)\big]\\ &\quad = \mathbb{P}[X_{-t} = x]\mu(x_0, \ldots, x_n)1_{\{t\leq n\}},\end{aligned}$$

so

$$\begin{aligned}\mathbb{P}[X_0 = y] &= \sum_{t=1}^{\infty}\sum_{\vec{x}} \mathbb{P}[X_{-t} = x]\mu(x_0, \ldots, x_n)1_{\{t\leq n\}}1_{\{x_t=y\}}\\ &= \mathbb{P}[X_{-t} = x]\sum_{t=1}^{\infty}\mathbb{P}[X_t = y,\ t \leq \tau_x^+].\end{aligned}$$

Summing over y, we see that $\mathbb{P}[X_{-t} = x] = \mathbb{E}_x[\tau_x^+]^{-1}$.

Invariant laws

We can turn this idea around and define

$$\pi(y) := \mathbb{E}_x[\tau_x^+]^{-1} \sum_{t=1}^{\infty} \mathbb{P}[X_t = y,\ t \leq \tau_x^+].$$

One can then check that this indeed defines an invariant law (see the book).

This proves existence of an invariant law. We postpone the proof of uniqueness till later.

The fact that there are no invariant laws on inessential classes follows from our earlier lemma, wich says that the Markov chain exits such classes in finite expected time. ∎

The Convergence Theorem

Theorem Let $(X_t)_{t\geq 0}$ be an irreducible aperiodic Markov chain with finite state space and let π be its invariant law. Then

$$\mathbb{P}_\mu[X_t = x] \underset{t\to\infty}{\longrightarrow} \pi(x) \qquad (x \in \mathcal{X}).$$

Remark The proof works for any invarant law π. Uniqueness of π (for aperiodic chains) then follows from the theorem.

Proof Let P be the transition kernel and $\mathcal{X}$ the state space.
Let $(X_t)_{t\geq 0}$ and $(Y_t)_{t\geq 0}$ be independent Markov chains with transiton kernel P and initial laws μ, ν.
Then $(X_t, Y_t)_{t\geq 0}$ is a Markov chain with transition kernel $\overline{P}(x, y; x', y') = P(x, x')P(y, y')$.
Since P is aperiodic, there exists t such that $P^t(x, y) > 0$ for all x, y and hence $\overline{P}^t(x, y; x', y') = P^t(x, x')P^t(y, y') > 0$ for all $(x, y), (x', y')$, proving that $\overline{P}$ is irreducible.

The Convergence Theorem

Define

$$\tau_{\mathrm{c}} := \inf\{t \geq 0 : X_t = Y_t\}.$$

Since $\overline{P}$ is irreducible, $\mathbb{E}[\tau_{\mathrm{c}}] < \infty$. Let

$$Y'_t := \begin{cases} Y_t & \text{if } t \leq \tau_{\mathrm{c}}, \\ X_t & \text{if } t \geq \tau_{\mathrm{c}}. \end{cases}$$

Then $(Y'_t)_{t\geq 0}$ is equal in law to $(Y_t)_{t\geq 0}$ and

$$\mathbb{P}_\mu[X_t = x] - \mathbb{P}_\nu[Y_t = x] \leq \mathbb{P}[X_t \neq Y'_t] \leq \mathbb{P}[t < \tau_{\mathrm{c}}] \underset{t\to\infty}{\longrightarrow} 0$$

for all $x \in \mathcal{X}$. In particular, if $\mu = \pi$ is any invariant law, then $\mathbb{P}_\pi[X_t = x] = \pi(x)$ and the claim follows. $\blacksquare$

Remark 1 The proof actually shows that

$$\|\mu P^t - \nu P^t\|_{\mathrm{TV}} := \sum_{x \in \mathcal{X}} \big|\mu P^t(x) - \nu P^t(x)\big| \leq \mathbb{P}[t < \tau_{\mathrm{c}}],$$

which by an earlier lemma goes to zero exponentially fast.

Remark 2 If P is a irreducible probability kernel, then

$$Q(x,y) := \tfrac{1}{2}P(x,y) + \tfrac{1}{2}1_{\{x=y\}}$$

is called the *lazy version* of P. Since Q is always aperiodic, it has a unique invariant law. But each invariant law π of P also solves $\pi Q = \pi(\frac{1}{2}P + \frac{1}{2}1) = \frac{1}{2}\pi P + \frac{1}{2}\pi = \pi$. Thus, P has a unique invariant law too.

Reversibility

Let $(X_t)_{t\in\mathbb{Z}}$ is a stationary Markov chain with transition kernel P and invariant law π.
Then $(X_{-t})_{t\in\mathbb{Z}}$ is a stationary Markov chain with transition kernel

$$\hat{P}(x,y) = \frac{\pi(y)P(y,x)}{\pi(x)}.$$

The chain $(X_t)_{t\in\mathbb{Z}}$ and the *reversed chain* $(X_{-t})_{t\in\mathbb{Z}}$ are equal in law if and only if *detailed balance* holds:

$$\pi(x)P(x,y) = \pi(y)P(y,x).$$

In general: $\pi(x)P(x,y) = \mathbb{P}[X_0 = x,\ X_1 = y] = \pi(y)\hat{P}(y,x)$.
A measure satisfying detailed balance is called a *reversible measure*.

Reversibility

Example Our earlier Markov chain $(Z_t, Z_{t+1}, Z_{t+2})_{t\geq 0}$ taking values in $\{0,1\}^3$ is not reversible.
In the forward time direction, $(z_1, z_2, z_3) \mapsto (z_2, z_3, 0)$ or $\mapsto (z_2, z_3, 1)$ with equal probabilities. The invariant law π is the uniform law on $\{0,1\}^3$ and the reversed chain makes the transitions $(z_1, z_2, z_3) \mapsto (0, z_1, z_2)$ or $\mapsto (1, z_1, z_2)$ with equal probabilities.

Harmonic functions

A function $h : \mathcal{X} \to \mathbb{R}$ satisfying $Ph = h$ is called *harmonic*.

Lemma A harmonic function is constant on each essential communicating class, and uniquely determined by its values on the essential communicating classes.

Proof Let h be harmonic and let C be an essential communicating class. Let $x \in C$ be such that $h(y) \leq h(x)$ for all $y \in C$. Then

$$h(x) = \sum_{y \in C} P(x,y)h(y) \leq h(x),$$

with equality if and only if $h(x) = h(y)$ for all y such that $P(x,y) > 0$. By induction, $h(x) = h(y)$ for all $y \in C$.

Harmonic functions

For any $x \in \mathcal{X}$,

$$h(x) = P^t h(x) = \mathbb{P}_x[h(X_t)] \qquad (t \geq 0).$$

Let $C_1, \ldots, C_n$ be the essential classes and let $h(x) = c_i$ for all $i \in C_i$.
Set $\tau := \inf\{t \geq 0 : X_t \in C_1 \cup \cdots \cup C_n\}$. Then

$$\begin{aligned} h(x) &= \mathbb{P}_x[h(X_t)] = \mathbb{P}_x[h(X_t) 1_{\{t<\tau\}}] + \sum_{i=1}^{n} c_i \mathbb{P}_x[X_t \in C_i] \\ &\underset{t\to\infty}{\longrightarrow} \sum_{i=1}^{n} c_i \mathbb{P}[X_\tau \in C_i]. \end{aligned}$$

∎

Harmonic functions

More precise lemma Let $C_1, \ldots, C_n$ be the essential classes. Then for each i, there exists a unique harmonic function h_i such that

$$h_i(x) = 1 \quad (x \in C_i) \quad \text{and} \quad h_i(x) = 0 \quad (x \in C_j,\ j \neq i).$$

This function is given by

$$h_i(x) = \mathbb{P}_x[X_\tau \in C_i] \qquad (x \in \mathcal{X}).$$

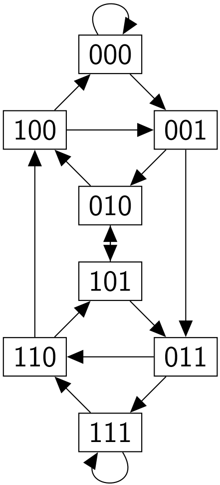
Moreover, each harmonic function is a linear combination of the functions $h_1, \ldots, h_n$.

Harmonic functions

Example Let $(Z_t)_{t\geq 0}$ be i.i.d. Bernoulli random variables with $\mathbb{P}[Z_t = 0] = \mathbb{P}[Z_t = 1] = \frac{1}{2}$ and let

$$\tau_{110} := \inf\big\{t \geq 0 : (Z_t, Z_{t+1} Z_{t+2}) = (1,1,0)\big\},$$
$$\tau_{010} := \inf\big\{t \geq 0 : (Z_t, Z_{t+1} Z_{t+2}) = (0,1,0)\big\}.$$

We can calculate $\mathbb{P}[\tau_{010} < \tau_{110}] = \pi h = 3/8$ where π is the uniform distribution on $\{0,1\}^3$ and h is the harmonic function on the right.



Harmonic functions

Example Let $(Z_t)_{t\geq 0}$ be i.i.d. Bernoulli random variables with $\mathbb{P}[Z_t = 0] = \mathbb{P}[Z_t = 1] = \frac{1}{2}$ and let

$$\tau_{110} := \inf\big\{t \geq 0 : (Z_t, Z_{t+1} Z_{t+2}) = (1,1,0)\big\},$$
$$\tau_{010} := \inf\big\{t \geq 0 : (Z_t, Z_{t+1} Z_{t+2}) = (0,1,0)\big\}.$$

We can calculate $\mathbb{P}[\tau_{010} < \tau_{110}] = \pi h = 3/8$ where π is the uniform distribution on $\{0,1\}^3$ and h is the harmonic function on the right.