**MASSACHUSETTS INSTITUTE OF TECHNOLOGY**

6.265/15.070J Lecture 9 — Mar 15, SP17

Lecturer: Yury Polyanskiy — Scribe notes by Themis Gouleakis

**Disclaimer:** *These notes have not been subjected to the usual scrutiny reserved for formal publications. They are posted to serve class purposes.*

# Introduction to Markov chains

**Content.**

1. Markov chain example and definitions
2. Strong Markov property
3. Existence and uniqueness of the stationary distribution
4. Reversibility of Markov chains

## 1 Markov chain example and definitions

### 1.1 Markov chain motivating example

We consider the following problem: We want to generate a bit string uniformly at random among all the $0-1$ bit strings of length $n$ that have no two adjacent 1's. Some approaches:

- simple (and slow): generate a random string uniformly and then check if it has the desired property (i.e rejection sampling). This is clearly an exponential algorithm in expectation, since there are only about $1.7^n$ (out of the $2^n$) strings of length $n$ which have this property.
- simple (and incorrect): generate first letter uniformly; subsequent letters are generated uniformly if previous one is $0$, or just frozen to $0$ if previous letter is 1. Equivalently: generate $N \gg n$ fair coin flips and delete adjacent ones. (Does not work because $P[0000] < P[01010]$.)
- tedious: compute the number of strings with first 0 and first 1 (generating functions help), then generate the first bit. Proceed recursively.

A much nicer/faster algorithm for sampling from this distribution turns out to be the following:

1. Initialize with a string of all $0$'s

2. **For** i=1 **to** k **do**: Pick a coordinate u.a.r. If the are no neighboring $0$'s resample it uniformly from $\{0, 1\}$.

In order to prove correctness of this algorithm, we need to show that as $k \to \infty$ the distribution of the output string approaches the desired one. For proving efficiency, we need to specify is how $k$ has to be in order to get a good enough approximation. This is an informal definition of the *mixing time* of a Markov chain.

The above example shows that one of the main motivations for studying Markov chains is that it gives us the ability to sample more efficiently from distributions with huge state spaces which wouldn't be possible to do with standard techniques like rejection sampling.

### 1.2 Definitions

**Definition 1.** *A sequence of random variables $(X_0, \ldots, X_N)$ (with $N = \infty$ permitted) is called* Markov process *with* transition kernel $P(a, b)$ *(which is a row stochastic matrix) and initial distribution $P_{X_0}$ if*

$$\mathbb{P}[X_0 = a_0, \ldots, X_n = a_n] = P_{X_0}[a_0] \prod_{i=1}^{n} P(a_{i-1}, a_i)$$

*for all $a^n \in \mathcal{X}^n$ and all $n \leq N$.*

Sometimes it is useful to consider Markov chains with state transitions that correspond to $t > 1$ transitions of a simpler Markov chain:

**Definition 2.** *Given a transition kernel $P(a, b)$, we define the corresponding* $t$-step transition kernel *to be:*

$$P_t(a, b) = \sum_{a_1, \ldots, a_{t-1}} P(a, a_1) \cdot P(a_1, a_2) \cdot \ldots \cdot P(a_{t-1}, b)$$

As a convention, a transition kernel $P$ acts on distributions on the left as follows:

$$\pi \to \pi P := \sum_b \pi(b) P(b, a)$$

and on functions on the right as follows:

$$f \to Pf(a) := \sum_b P(a, b) f(b) = \mathbb{E}[f(X_{i+1}) | X_i = a]$$

**Definition 3.** *We will say that a Markov chain that has a transition kernel* $P(a, b)$ *is* irreducible *if*

$$\forall a, b \in \mathcal{X} : \exists t > 0 : P_t(a, b) > 0$$

In other words, a Markov chain is irreducible if and only if every state is reachable from every other state. It is also true that every Markov chain is irreducible for a subset of its state space. The states that belong to that subset are called *recurrent states*. All other states of the Markov chain are called *transient*. If the is only one recurrent state, then this state is called *absorbing state*.

From the above, it is clear that if we sample a state randomly accodring to some distribution $d$ and perform one transition of the Markov chain, the new state will be distributed according to the distribution $d' = dP$. If we continue making transition long enough, we might see convergence to one of the fixed points of this mapping defined as follows:

**Definition 4.** *A distribution* $\pi$ *over* $\mathcal{X}$ *will be called* stationary distribution *if and only if the following holds:* $\pi = \pi P$.

The following definition and proposition will be useful for talking about reversibility of Markov chains:

**Definition 5.** *If* $\pi$ *is a stationary distribution of a Markov chain with transition kernel* $P$*, then* $\tilde{P}(a, b) = \frac{P(b,a)}{\pi(a)}\pi(b)$ *is called* reverse transition kernel.

**Proposition 1.** *If* $(X_0, \ldots, X_N)$ *is a Markov chain started from stationary distribution, then* $(X_n, \ldots, X_0)$ *is also Markov (for any* $n$*) with transition kernel* $\tilde{P}$ *and started from the same stationary distribution.*

Even though the Markon chain visits recurrent states infinitely often (if we run it indefinitely), it is possible that visits to a particular vertex can occur only periodically.

**Definition 6.** *Let* $x \in \mathcal{X}$ *be a state of the Markov chain. The* period *of* $x$ *is defined as* $gcd\{t > 0 : P_t(x, x) > 0\}$.

**Proposition 2.** *If* $P$ *is irreducible, then all periods are equal. An irreducible Markov chain is called* aperiodic *iff all periods are equal to* $1$.

Finally, we define the *hitting time* to be the number of steps until the Markov chain reaches a subset $B$ of the state space.

**Definition 7.** *Let $B \subseteq \mathcal{X}$ be a set of states of the Markov chain. Then the following quantity is called* hitting time *of $B$:*

$$\tau_B = inf\{t \geq 0 : \mathcal{X}_t \in B\}$$

$$\tau_B^+ = inf\{t > 0 : \mathcal{X}_t \in B\}$$

## 2 Strong Markov property

All Markov chains have the "Markov property" by definition, which can be stated as follows:

$$\mathbb{P}[X_{i+1}|X_i, X_{i-1}, \ldots, X_0] = \mathbb{P}[X_{i+1}|X_i]$$

We will get a stronger peoperty if we allow $k$ to be a stopping time random variable.

**Theorem 1.** *Suppose $\{X_i\}$ is a Markov chain described by a transition kernel $P$ and $\tau$ is a stopping time such that $\tau < \infty$ almost surely. Then $(Y_0, \ldots, Y_n)$, where $Y_i := X_{\tau+i}$, is a Markov chain with the same transition kernel.*

*Proof.* We want to show that $(Y_0, \ldots, Y_n)$ has the Markov property. Thus, we want to compute:

$$\begin{aligned}
\mathbb{P}[Y_0 = a_0, Y_1 = a_1, \ldots, Y_n = a_n] &= \sum_{r=0}^{\infty} \mathbb{P}[Y_0 = a_0, Y_1 = a_1, \ldots, Y_n = a_n, \tau = r] \\
&= \sum_{r=0}^{\infty} \mathbb{P}[X_\tau = a_0, \ldots, X_{\tau+n} = a_n, \tau = r] \\
&= \sum_{r=0}^{\infty} \mathbb{P}[X_\tau = a_0, \tau = r] \prod_{i=1}^{n} P(a_{i-1}, a_i) \\
&= \mathbb{P}[X_\tau = a_0] \prod_{i=1}^{n} P(a_{i-1}, a_i) \\
&= \mathbb{P}[Y_0 = a_0] \prod_{i=1}^{n} P(a_{i-1}, a_i)
\end{aligned}$$

as required. □

An example of a Markov chain that has the strong Markov property is the symmetric random walk on $\mathbb{Z}$ and the following distribution is an example of a Markov chain that has the Markov property but not the strong Markov property: $X_t = \begin{cases} 0 & \text{for } t \leq T, T \sim Exp(1) \\ t - T & \text{for } t > T \end{cases}$ If we consider the stopping time $\tau = \inf\{t > 0 : X_t > 0\}$, then we know that $X_{\tau+i} = i + 1$ deterministically. Therefore, the Markov chain does not have the strong Markov property.

## 3 Existence and uniqueness of the stationary distribution

The following theorem establishes the existence and the uniqueness of the stationary distribution.

**Theorem 2.** *Let $P$ be the transition kernel of some irreducible Markov chain on a finite domain $\mathcal{X}$. Then, the Markov chain has a unique stationary distribution $\pi$ and $\pi(a) = \frac{1}{\mathbb{E}^\alpha[\tau_\alpha^+]}$*

*Proof.*
**Uniqueness:** If $\pi$ is a stationary distribution, then we have:

$$\begin{aligned}\pi(\alpha)\mathbb{E}[\tau_\alpha^+] = \sum_{t\geq 0} \pi(\alpha)\mathbb{P}^\alpha[\tau_\alpha^+ > t] &= \sum_{t\geq 0} \mathbb{P}^\pi[X_0 = \alpha, X_1 \neq \alpha, \dots, X_t \neq \alpha] \\ &= \sum_{t\geq 0} \mathbb{P}^\pi[\tilde{X}_t = \alpha, \tilde{X}_{t-1} \neq \alpha, \dots, X_0 \neq \alpha] \\ &= \mathbb{P}[\tilde{X} \text{ ever hits } \alpha] = 1\end{aligned}$$

where $(\tilde{X}_0, \tilde{X}_1, \dots)$ are samples from the reverse Markov chain with kernel $\tilde{P}$, and the last equality follows from the irreducibility of $P$.

**Lemma 3.** *If $P$ is the kernel of an irreducible Markov chain, then $\mathbb{E}^\alpha[\tau_b] < \infty$.*

**Existence:** Let

$$N(b) := \sum_{0\leq t<\tau_\alpha^+} I\{X_{t+1} = b\} = \sum_{0\leq t<\tau_\alpha^+} \sum_c I\{X_t = c, X_{t+1} = b\}$$

So, the expected number of visits to $b$ before returning to $\alpha$ is the following:

$$\mathbb{E}[N(b)] = \sum_c \mathbb{E}[N(c)]P(c, b)$$

That is, $\pi(b) \triangleq \frac{\mathbb{E}[N(b)]}{\sum_{b'} \mathbb{E}[N(b')]}$ is a stationary distribution. □

The first part of the proof is just a specialization of the following surprising property:

**Lemma 4** (Kac Lemma)**.** *For any stationary ergodic process* $\{X_0, \ldots\}$ *we have*

$$\mathbb{E}[\tau_a^+ | X_0 = a] = \frac{1}{\mathbb{P}[X_0 = a]}$$

## 4 Reversibility of Markov chains

Intuitively, a Markov chain is called *time-reversible* if for each pair of states $i, j$ the long-run rate at which the chain makes a transition from state $i$ to state $j$ equals the long-run rate at which the chain makes a transition from state $j$ to state $i$: $\pi_i P(i,j) = \pi_j P(j,i)$. The formal definition follows:

**Definition 8.** *A Markov chain is called* reversible *at stationary distribution* $\pi$ *if* $\tilde{P} = P$*, in which case* $\pi$ *is called reversing measure.*

**Proposition 3.** *A Markov chain with transition kernel* $P$ *is reversible iff there exists a solution to the detailed-balance equation:*

$$\pi(a)P(a,b) = \pi(b)P(b,a)\,, \tag{1}$$

*in which case* $\pi$ *is the reversing measure (and hence a stationary distribution of* $P$*).*

**Example 1:**(non-reversible)
An example of a non reversible Markov chain is a cycle where all clockwise transitions have probability $2/3$ and all counterclockwise transitions have probability $1/3$. Clearly that stationary distribution is uniform due to symmetry. So, using proposition 3 we see that there is no solution since $\pi(a) = \pi(b)$ and $P(a,b) \neq P(b,a)$ for all $a, b$.

**Example 2:**(reversible)
An example of a reversible Markov chain is a random walk on a weighted graph where the transition probabilties are defined as follows:

$$P(i,j) := \frac{w_{ij}}{\sum_l w_{il}}$$

The stationary distribution of this Markov chain is:

$$\pi(i) = \frac{1}{Z} \sum_i w_{il}$$

where $Z$ is just the appropriate normalization factor for the probabilities to sum to 1.

**Proposition 4.** *All reversible Markov chains are random walks on a undirected weighted graph.*

*Proof.* (sketch) Setting the weights as follows:

$$w_{ij} := \pi(i)P(i,j) = \pi(j)P(j,i)$$

one can verify that

$$P(i,j) := \frac{w_{ij}}{\sum_l w_{il}}$$

corresponds to the stationary distribution $\pi(i) = \frac{1}{Z}\sum_j w_{ij}$.

Also, note that those weigths could not be defined if the Markov cahin was not reversible. □

Since in an undirected graph the smallest cycle that contains a particular vertex is 2, proposition 4 has the following corollary:

**Corollary 5.** *Reversible Markov chains can have only period* 1 *or* 2.

In particular, for irreducible Markov chains the period is 2 if and only if the graph is bipartite. Otherwise, the graph has at one odd cycle and is strongly connected due to irreduciility. Therefore, every vertex is in both a length 2 cycle and in an odd length cycle, implying that the gcd is 1.

**Proposition 5.** *Birth-death chains are Markov chains on* $\{0, 1, 2, 3, \ldots, n\}$ *such that* $P(a, b) = 0$ *for any* $|a - b| > 1$. *All birth-death Markov chains are reversible.*

*Proof.* Just note that detailed balance equation (1) is always solvable recursively: $\pi(1)$ is written as a function of $\pi(0)$, $\pi(2)$ as a function of $\pi(1)$, etc. Then normalize. □