STATS 310B: Theory of Probability II — Winter 2016/17

Lecture 14: February 23

Lecturer: Sourav Chatterjee — *Scribes: Kenneth Tay*

14.1 Markov Chains

Definition 14.1 *Let $(\Omega, \mathcal{F}, P)$ be a probability space, $\{\mathcal{F}_n\}_{n\geq 0}$ a filtration, $\{X_n\}_{n\geq 0}$ an adapted sequence of random variables taking values in some measurable space $(S, \mathcal{S})$.*

*$\{X_n\}_{n\geq 0}$ is called a **Markov chain** if for all n and all $B \in \mathcal{S}$,*

$$P(X_{n+1} \in B \mid \mathcal{F}_n) = P(X_{n+1} \in B \mid X_n).$$

Example

$X_n = \sum_{i=0}^{n} Y_i$, where Y_i are independent. For this process, we have

$$\begin{aligned} P(X_{n+1} \in B \mid \mathcal{F}_n) &= P(X_n + Y_{n+1} \in B \mid \mathcal{F}_n) \\ &= P(Y_{n+1} \in B - X_n \mid \mathcal{F}_n) \\ &= \mu_{n+1}(B - X_n). \qquad (\mu_{n+1} \text{ is the law of } Y_{n+1}) \end{aligned}$$

(You should check the last step.) A similar calculation shows that $P(X_{n+1} \in B \mid X_n) = \mu_{n+1}(B - X_n)$ as well.

For simplicity, we will henceforth work with countable state space S. With this assumption, quantities of the form $P(X_0 = x_0, \ldots, X_n = x_n)$ makes sense. The definition of a Markov chain implies that

$$P(X_{n+1} = x_{n+1} \mid X_0 = x_0, \ldots, X_n = x_n) = P(X_{n+1} = x_{n+1} \mid X_n = x_n).$$

Definition 14.2 ***Transition probabilities*** *are defined as*

$$p_n(x, y) = P(X_{n+1} = y \mid X_n = x).$$

The transition probabilities determine the behavior of the Markov chain.

Theorem 14.3 (Chapman-Kolmogorov Equations) *Let $P_n = \Big(p_k(x,y)\Big)_{x,y\in S}$, and $P^{(n)} = P_0 P_1 \ldots P_{n-1}$. Then*

$$P(X_n = y \mid X_0 = x) = P^{(n)}(x, y).$$

Proof:

$$\begin{aligned}
P(X_n = y \mid X_0 = x) &= \sum_{x_1,\dots,x_{n-1}} P(X_n = y, X_{n-1} = x_{n-1}, \dots, X_1 = x_1 \mid X_0 = x) \\
&= \sum_{x_1,\dots,x_{n-1}} P(X_n = y \mid X_0 = x, X_1 = x_1, \dots, X_{n-1} = x_{n-1}) \\
&\qquad \cdot P(X_{n-1} = x_{n-1} \mid X_0 = x, \dots, X_{n-2} = x_{n-2}) \cdot \ldots \cdot P(X_1 = x_1 \mid X_0 = x) \\
&= \sum_{x_1,\dots,x_{n-1}} p_{n-1}(x_{n-1}, y) p_{n-2}(x_{n-2}, x_{n-1}) \dots p_0(x_2, x_1) \\
&= P^{(n)}(x, y).
\end{aligned}$$

∎

Definition 14.4 *A Markov chain is **time-homogeneous** if the transition probabilities p_n are the same for all n.*

*In this case, we simply write p instead of p_n, and we call $P = \Big(p(x,y)\Big)_{x,y\in S}$ the **transition matrix**.*

In addition, we have $P^{(n)} = P^n$ for all n.

Example: Sums of i.i.d. random variables.

In this class, we will generally work only with time-homogeneous Markov chains.

14.1.1 Markov Property and Strong Markov Property

Let $\{X_n\}_{n\geq 0}$ be a time-homogeneous Markov chain.

Proposition 14.5 *$\{X_n\}$ has the **Markov property**, i.e. given a function $f : S \times S \times \ldots \to \mathbb{R}$ which is measurable w.r.t. the product σ-algebra, let $g(x) = \mathbb{E}[f(X_0, X_1, X_2, \dots) \mid X_0 = x] =: \mathbb{E}_x[f(X_0, X_1, X_2, \dots)]$. Then, for any n,*

$$\mathbb{E}[f(X_n, X_{n+1}, X_{n+2}, \dots) \mid X_n = x] = g(x).$$

Proof: The result is immediate because the joint law of $(X_0, X_1, \dots)$ given $X_0 = x$ is the same as the joint law of $(X_n, X_{n+1}, \dots)$ given $X_n = x$.

∎

(**Note:** We can show that $g(X_n) = \mathbb{E}[f(X_n, X_{n+1}, \dots) \mid \mathcal{F}_n]$.)

Proposition 14.6 *$\{X_n\}$ has the **Strong Markov property**, i.e. with the same set-up above, let T is a stopping time w.r.t. $\{\mathcal{F}_n\}_{n\geq 0}$. Then, on the set $\{T < \infty\}$,*

$$\mathbb{E}[f(X_T, X_{T+1}, \dots) \mid \mathcal{F}_T] = g(X_T).$$

Proof: We show that $g(X_T)$ satisfies the definition of conditional expectation for $\mathbb{E}[f(X_T, X_{T+1}, \dots) \mid \mathcal{F}_T]$.

Take any $A \in \mathcal{F}_T$ and $A \subseteq \{T < \infty\}$, and any n. Then $A \cap \{T = n\}$ belongs to $\mathcal{F}_n$, and so we have

$$\begin{aligned}
\int_{A\cap\{T=n\}} g(X_T)dP &= \int_{A\cap\{T=n\}} g(X_n)dP \\
&= \int_{A\cap\{T=n\}} \mathbb{E}[f(X_n, X_{n+1}, \dots) \mid \mathcal{F}_n]dP \\
&= \int_{A\cap\{T=n\}} f(X_n, X_{n+1}, \dots)dP && \text{(since } A\cap\{T=n\} \in \mathcal{F}_n\text{)} \\
&= \int_{A\cap\{T=n\}} f(X_T, X_{T+1}, \dots)dP.
\end{aligned}$$

We get our desired result by summing both sides over n.

■

14.1.2 Hitting Times, Recurrence and Transience

Let $\{X_n\}_{n\geq 0}$ be a time-homogeneous Markov chain taking values on a countable state space S.

Definition 14.7 *For $x \in S$, the **first hitting time** of x is $T_x := \inf\{n \geq 1 : X_n = x\}$. (Note: Time 0 doesn't count.)*

Definition 14.8 *Let $\rho_{xy} := P(T_y < \infty \mid X_0 = x)$. In particular, ρ_{xx} is the probability of ever returning to x given that the chain starts at x.*

*We say that a state x is **recurrent** if $\rho_{xx} = 1$, and is **transient** otherwise.*

Let $N(x) = \sum_{n=1}^{\infty} 1_{\{X_n = x\}}$, i.e. the number of visits to x (except time 0).

Theorem 14.9 *The following are equivalent:*

(a) x is recurrent.

(b) $\mathbb{E}_x N(x) = \infty$, where $\mathbb{E}_x$ means $\mathbb{E}[\cdot \mid X_0 = x]$.

(c) $P_x(N(x) = \infty) = 1$.

This theorem is useful because, by the Monotone Convergence Theorem,

$$\mathbb{E}_x N(x) = \sum_{n=1}^{\infty} P(X_n = x \mid X_0 = x).$$

The convergence or divergence of the sum on the RHS determines whether a state is recurrent or not.

14.2 Aside: Example of why undergraduate definitions of conditional expectation don't work

Take the unit disk and choose (X, Y) uniformly from the disk. **What is the distribution of X given $Y = 0$?**

We can compute it in 2 ways:

- The joint density of X and Y is given by

$$f(x,y) = \begin{cases} \frac{1}{\pi} & \text{if } (x,y) \text{ belongs to the disk,} \\ 0 & \text{otherwise.} \end{cases}$$

  Then the conditional density of X given $Y = 0$ is

$$f_{X|Y=0}(x) = \frac{f(x,y)}{f_Y(0)} = \frac{1}{2}$$

  for all $x \in [-1, 1]$.

- Suppose we do a change of variables to polar coordinates: $X = R\cos\Theta$, $Y = R\sin\Theta$, where $0 \le R \le 1$, $0 \le \Theta \le 2\pi$. The joint density of (R, Θ) is

$$f(r,\theta) = \begin{cases} \frac{r}{\pi} & \text{if } 0 \le r \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

  Thus, (R, Θ) are independent. The event$\{Y = 0\} = \{\Theta = 0 \text{ or } \pi\}$, and by symmetry, $\Theta = 0$ or π each with probability $\frac{1}{2}$. But R is independent of Θ, so given $\{\Theta = 0 \text{ or } \pi\}$, R is still distributed as a random variable with density $2r$ on $[0, 1]$. Since $X = R\cos\Theta$, given $Y = 0$, X has density $|x|$ on $[-1, 1]$.

That's why the undergraduate definitions of conditional probability can be problematic!