STA4502: Topics in Stochastic Processes Winter 2018

Lecture 1: Introduction — March 1, 2018

Lecturer: Jeffrey Rosenthal *Scribe: Mufan (Bill) Li*

1 Introduction to Markov Chains

Before we define a Markov chain, we introduce the following notations and definitions.

Definition 1. We denote $(\mathcal{X}, \mathcal{F})$ our (measurable) **state space**, equipped with appropriate σ-algebra $\mathcal{F}$. Here a state space is called **discrete** if the cardinality of $\mathcal{X}$ is finite or countable, otherwise it's called **continuous**.

Here's a diagram with some examples

$$\text{state space}(\mathcal{X}, \mathcal{F}) \begin{cases} \text{discrete:} \begin{cases} \text{finite:} & \text{e.g. } \mathcal{X} = \{1,2,3,\dots,n\}, \mathcal{F} = \mathcal{P}(\mathcal{X}) \text{ (the power set)} \\ \text{countable:} & \text{e.g. } \mathcal{X} = \mathbb{N} = \{0,1,2,\dots\}, \mathcal{F} = \mathcal{P}(\mathcal{X}) \end{cases} \\ \text{continuous: e.g. } \mathcal{X} = \mathbb{R}^d, \mathcal{F} = \mathcal{B}(\mathbb{R}^d) \text{ (the Borel sets).} \end{cases}$$

Definition 2. $P : \mathcal{X} \times \mathcal{F} \to [0,1]$ is a **transition probability** if $\forall x \in \mathcal{X}$, $P(x, \cdot) : \mathcal{F} \to [0,1]$ is a probability measure on $(\mathcal{X}, \mathcal{F})$.

Remark 3. Observe that for a discrete $\mathcal{X}$, $P(x,y)$ is well defined $\forall x, y \in \mathcal{X}$, i.e. it's the probability of going from point x to point y. For a continuous $\mathcal{X}$, we need to use measurable sets $A \in \mathcal{F}$ to talk about transition probability $P(x, A)$ from x to A.

Finally we can define a Markov chain.

Definition 4. We call a sequence of random variables $\{X_k\}_{k=0}^{\infty}$ taking values in $\mathcal{X}$ a **Markov chain** if $\mathbb{P}[X_{k+1} \in A | X_k] = P(X_k, A), \forall k \in \mathbb{N}, A \in \mathcal{F}$.

At the same time, we would like to define the following distributions on $\mathcal{X}$.

Definition 5. We call $\nu = \mathcal{L}(X_0)$ the **initial disitribution**, and denote $\mu_k = \mathcal{L}(X_k)$. A distribution π is called a **stationary distribution** if $\pi(A) = \int_{\mathcal{X}} P(x, A) d\pi(x)$.

In other words, if the Markov chain starts in a stationary distribution ($\nu = \pi$), it will remain in stationarity ($\mu_k = \pi, \forall k \in \mathbb{N}$).

The main goal of this course is study whether or not we have convergence of $\mu_k \to \pi$ in some sense, and if so quantify the "rates" of this convergence. To this end, we will introduce several defintions and basic conditions to guarantee convergence.

2 Convergence Conditions

We start with a couple of definitions.

Definition 6. A dicrete Markov chain is **irreducible** if

$$\forall x, y \in \mathcal{X}, \mathbb{P}[X_k = y \text{ eventually } | X_0 = x] > 0.$$

Equivalently, we can say $\exists m \in \mathbb{N} : P^m(x, y) > 0$, where $P^m(x, y)$ is the transition probability after m steps.

In general, we say Markov chain is ϕ-**irreducible** if $\exists$ non-zero σ-finite measure ϕ on $(\mathcal{X}, \mathcal{F}) : \forall x \in \mathcal{X}, A \in \mathcal{F}$ with $\phi(A) > 0$, we have that $\mathbb{P}[X_k \in A \text{ eventually } | X_0 = x] > 0$.

Remark 7. The general ϕ-irreducibility is not equivalent to the discrete irreducibility. Consider for example ϕ that concentrates only on a single point $x \in \mathcal{X}$, then ϕ-irreducibility only requires $\forall y \in \mathcal{X}, \exists m \in \mathbb{N} : P^m(y, x) > 0$.

Definition 8. For a discrete irreducible Markov chain, a point $x \in \mathcal{X}$ is said to be **aperiodic** if $\gcd(\{n, P^n(x, x) > 0\}) = 1$. The Markov chain is aperiodic if every point is aperiodic.

A general Markov chain with stationary distribution π is **aperiodic** if there does not exist $d \geq 2$ and a partition of size $d + 1$ such that $\mathcal{X} = \left(\bigsqcup_{i=1}^d \mathcal{X}_i\right) \bigsqcup N$, where $\bigsqcup$ denotes disjoint union, N is a π-null set, and for π-a.e. $x \in \mathcal{X}_i, P(x, \mathcal{X}_{i+1}) = 1$, except for π-a.e. $x \in \mathcal{X}_d, P(x, \mathcal{X}_1) = 1$.

At this point, we can state our first theorem, the conditions to guarantee convergence.

Theorem 9. If a Markov chain is irreducible (or ϕ-irreducible for the general case), aperiodic, and have a stationary distribution π, then we have for a discrete Markov chain

$$\forall v \text{ initial distribution}, \forall y \in \mathcal{X}, \lim_{k\to\infty} \mu_k(y) = \pi(y),$$

or for a general Markov chain

$$\text{for } \pi - \text{a.e. } x \in \mathcal{X}, \lim_{k\to\infty} \sup_{A\in\mathcal{F}} \left| P^k(x, A) - \pi(A) \right| = 0.$$

Remark 10. For the general chain, this type of convergence is stronger than the typical weak convergence (in distribution), it is known as **convergence in total variation**. The name refers to the total variation distance defined by

$$\mathrm{TV}(\mu_k, \pi) := \sup_{A\in\mathcal{F}} |\mu_k(A) - \pi(A)|.$$

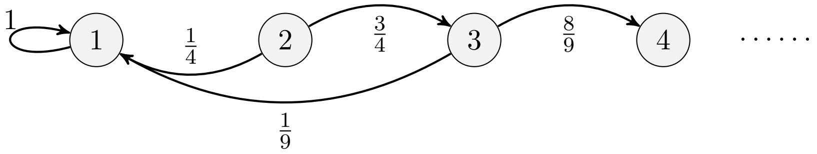
In the discrete case, we also have the following identity

$$\mathrm{TV}(\mu_k, \pi) = \frac{1}{2} \sum_{y\in\mathcal{X}} |\mu_k(y) - \pi(y)| = \sum_{y\in\mathcal{X}:\mu_k(y)>\pi(y)} |\mu_k(y) - \pi(y)|.$$

Therefore we have that for the discrete case, weak convergence implies convergence in total variation.

However, in general this implication is false. Consider the following counter example (by Jeffrey Negrea). Let $\mathcal{X} = [0, 1], \pi = \delta_0$, a point mass at $x = 0$. Define the transition probability as $P(x, x/2) = 1$. Then $\forall x \in \mathcal{X}, \nu = \delta_x$, we have $\mu_k \xrightarrow{d} \delta_0$. However the total variation distance is always 1 as $|\pi(0) - \mu_k(0)| = 1, \forall k$.

Example 11. Let $\mathcal{X} = \{1, 2, \ldots\}, \pi = \delta_1, P(1,1) = 1$. For $n \geq 2$, let $P(n, n+1) = 1 - 1/n^2$, $P(n,1) = 1/n^2$. See diagram below for a few sample points.



Observe in this case, the chain is not irreducible in the discrete definition, however it is ϕ-irreducible when $\phi = \delta_1$. Similarly, this Markov chain is aperiodic in the general sense, since the only possible node to return to has a period of 1.

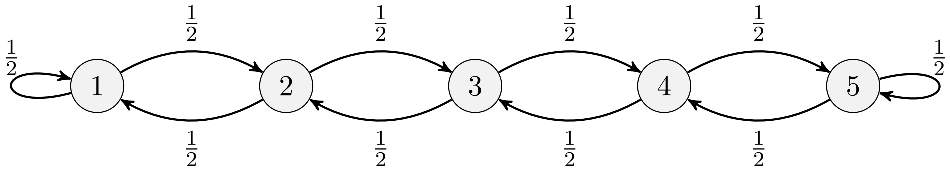
Here we can check for which $x \in \mathcal{X}$ we have convergence to stationary distribution.

$x = 1 \implies \mu_k = \delta_1 \, \forall k$. ✓

$x \geq 2$, in this case we have that $1/n^2$ is summable, which implies $\prod_{n=1}^{\infty}(1 - 1/n^2) > 0$. In other words, there is a positve probability of $X_k \to \infty$. ✗

This implies we have $\mu_k \xrightarrow{\text{TV}} \pi$, but only for π-a.e. $x \in \mathcal{X}$ starting points, which is only $x = 1$.

Example 12. Let $\mathcal{X} = \{1, 2, 3, 4, 5\}$, and define a symmetric random walk, i.e. $P(x, x+1) = P(x, x-1) = 1/2$, except at the ends, we have $P(1,1) = P(5,5) = 1/2$ instead. See diagram below.



This chain is clearly irreducible. It is also aperiodic since for every path $x \to y$, we can stop at 1 or 5 for one additional step, making the gcd 1.

The stationary distribution π is uniform since the chain is reversible, i.e. $P(x,y) = P(y,x)$.

Since all the conditions are satisfied we have

$$\lim_{k \to \infty} \mathbb{P}[X_k = x] = \pi(x) = \frac{1}{5}, \forall x \in \mathcal{X}.$$

Goal 13. Find $k^* \in \mathbb{N}$ such that

$$\sup_{A \in \mathcal{F}} \left| P^{k^*}(x, A) - \pi(A) \right| < 0.01$$

This is called the **quantitative rate of convergence**. Here we remark that finding one particular k^* is already difficult, therefore we are less interested in find the minimal k^*.

To this goal, we will introduce the coupling technique.

3 The Coupling Inequality

If X, Y are jointly defined random variables, then we can bound the total variation distance by the following steps

$$\begin{aligned} \|\mathcal{L}(X) - \mathcal{L}(Y)\|_{\mathrm{TV}} &= \sup_{A \in \mathcal{F}} |\mathbb{P}(X \in A) - \mathbb{P}(Y \in A)| \\ \text{partition each event } \ldots &= \sup_{A \in \mathcal{F}} |\mathbb{P}(X \in A, X = Y) + \mathbb{P}(X \in A, X \neq Y) \\ &\quad - \mathbb{P}(Y \in A, X = Y) - \mathbb{P}(Y \in A, X \neq Y)| \\ \{X \in A, X = Y\} = \{Y \in A, X = Y\} \implies &= \sup_{A \in \mathcal{F}} |\mathbb{P}(X \in A, X \neq Y) - \mathbb{P}(Y \in A, X \neq Y)| \\ &\leq \mathbb{P}(X \neq Y), \end{aligned}$$

where a factor of 2 is not required in the last step since both probabilities are non-negative.

Example 14. (Apply to Markov Chains) Here we start with a Markov chain $\{X_k\}_{k=0}^{\infty}$, and we make a copy of it denoted $\{Y_k\}_{k=0}^{\infty}$, with the joint distribution specified later.

Usually (although not exclusively), we will let $Y_0 \sim \pi$, i.e. start in the stationary distribution, therefore $Y_k \sim \pi$, i.e. remains in stationarity. This implies

$$\|\mu_k - \pi\|_{\mathrm{TV}} = \|\mathcal{L}(X_k) - \mathcal{L}(Y_k)\|_{\mathrm{TV}} \leq \mathcal{P}(X_k \neq Y_k).$$

Proof intuition: here we let X_k, Y_k move together, i.e. if $X_{k+1} = X_k \pm 1$, then we also have $Y_{k+1} = Y_k \pm 1$; while at the end points, one of the chains must remain in the same node, hence reducing the "distance" by 1. Eventually, the two chains will "converge" to the same value.

Challenge 15. Use the coupling technique above to find k^* such that

$$\sup_{A \in \mathcal{F}} \left| P^{k^*}(x, A) - \pi(A) \right| < 0.01$$