Note 11. Discrete-time Markov chains I

Alex Fu

Fall 2022

1 Introduction to random processes

Now having seen the foundations of probability theory — probability spaces, random variables and distributions, expectation and related functions — we can finally introduce the second part of this course's title, *random processes*.

Definition 1 (Random process).

A **random process** or *stochastic process* is an indexed family of random variables. Formally, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a common probability space, let (S, Σ) be a measurable space of values, and let T be an indexing set. Then a T-indexed S-valued random process is a family

$$\{X_t : \Omega \to S\}_{t \in T}\,.$$

T is commonly $\mathbb{N}$ to represent *discrete time* or $\mathbb{R}_{\geq 0}$ for *continuous time*.

The definition above is incredibly general, so virtually every natural phenomenon or system with randomness can be modelled by some random process. *Independent and identically distributed* collections of random variables are the simplest example of random processes, but these are far less interesting than processes that introduce some dependence.

Markov processes describe the simplest form of dependence: $* \to * \to \cdots$, in which a state depends only on its previous state. They thus describe systems and situations that are "predictably random," encompassing random walks, Brownian motion, and Poisson processes. We will focus mostly on Markov processes, giving a short list of other common random processes* below.

- *Queueing networks* describe random arrivals to and departures from one or more queues.
- *Martingales* describe the winnings of betting on a game fair in expectation.
- *Counting processes* describe the number of occurrences, e.g. arrivals, over time.
- *Branching processes* describe the size of a population through reproduction.

2 Introduction to dynamical systems

A **system** or *dynamical system* is, roughly speaking, something that evolves over time. Much of the terminology for random processes is shared with dynamical systems.

Definition 2 (Terminology of dynamical systems).

a. A system has **state**, a complete characterization of the system at any particular moment in time, and **dynamics** or *transitions*, rules that describe how the system changes from one state to a next state over time.

b. The **state space** S of a system is the collection of all possible states it can take on. We assume that S is countable, so without loss of generality $S \subseteq \mathbb{N}$.

c. A **discrete-time** system has time steps $n \in \mathbb{N}$, and a **continuous-time** system has times $t \in \mathbb{R}_{\geq 0}$. In both cases, $x_0 \in S$ is the **initial state** or *starting state*.

d. The **history** or *trajectory* of a system at time t is the collection of states $(x_s)_{s<t}$. We may also refer to **past** states, the **present** state, or **future** states. In discrete-time systems, x_{n-1} and x_{n+1} are the *immediate past* and *immediate future* respectively.

e. A system is **time-homogeneous** or *time-invariant* if its dynamics do not change over time. We will assume that systems are time-homogeneous throughout.

f. A system is **stationary**, *at equilibrium*, *in steady state*, or simply *invariant* if its state does not change over time. Stationarity will be a major behavior of interest for us.

g. A **deterministic** system is one whose behavior is fully determined given a starting state and the dynamics. A non-deterministic system is **random** or *stochastic*.

Characterizations of deterministic discrete-time systems lead nicely into those of random systems.

Example 1 (Dynamics function).

Let $x_0 \in S$ be a starting state, let $f\colon S \to S$ be a **dynamics function**, and set $x_{n+1} = f(x_n)$ for every $n \geq 1$. If the state space is $S = \mathbb{R}^d$ and the dynamics are a *linear* transformation, then the system is also given the matrix-vector equation $x_{n+1} = Ax_n$.

The k*-step evolution* is given by $x_{n+k} = (f \circ \overset{k}{\cdots} \circ f)(x_n)$, or more familiarly $x_{n+k} = A^k x_n$. The steady-state behavior is given by any fixed points $f(x) = x$, or any eigenvectors $v = Av$ with eigenvalue 1. Note that x_{n+1} is determined by x_n regardless of $x^{(n-1)} = (x_0, \ldots, x_{n-1})$.

Example 2 (Particle on a graph).

Let the state space S also be the vertex set of a graph, where an edge (i, j) exists iff $j = f(i)$ is a transition of the system. Equivalently, let the edge weight of (i, j) be $\mathbb{1}_{\{j=f(i)\}}$ in $\{0, 1\}$.

Then the present state is the position of a particle which starts at x_0 and jumps to the vertex $x_{n+1} = f(x_n)$ at every time step $n \geq 1$.

Here, the trajectory is the path or sequence of edges taken by the particle, and steady-state behavior is when the particle remains stationary at one state.

3 Four characterizations of Markov chains

3.1 Sequence of random variables

Definition 3 (Discrete-time Markov chain; Markov property).

A **discrete-time Markov chain** (DTMC), or simply *Markov chain* (MC), is a sequence of S-valued random variables $(X_n)_{n \in \mathbb{N}}$ that satisfies the **Markov property**:

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

for every $n \in \mathbb{N}$ and every *feasible* sequence of states $(x_0, \ldots, x_n)$ with nonzero probability.

The Markov property is often intuitively described as

"Given the present, the future is unaffected by knowledge about the past."

Or, "the future is dependent only on the present." Take care not to overextend this intuition — it is a statement of **conditional** independence, not independence.

The Markov property is sometimes also called *memorylessness*, as the system has no "memory" of its history when transitioning to future states. We will later find interesting connections between the Markov property and the memorylessness of the Geometric and Exponential distributions.

Definition 4 (Time homogeneity; transition probabilities).

A Markov chain is **time-homogenous**, *time-invariant*, or *temporally homogeneous* if for every time step $n \in \mathbb{N}$, time shift $k \in \mathbb{N}$, and states $i, j \in S$,

$$\mathbb{P}(X_{n+1} = j \mid X_n = i) = \mathbb{P}(X_{n+k+1} = j \mid X_{n+k} = i).$$

An apparently weaker but equivalent condition is

$$\mathbb{P}(X_{n+1} = j \mid X_n = i) = \mathbb{P}(X_{n+2} = j \mid X_{n+1} = i).$$

The **transition probabilities** $p_{i,j} = p(i,j) := \mathbb{P}(X_{n+1} = j \mid X_n = i)$ thus well-defined as quantities independent of time. We can also define the k-**step** transition probabilities

$$p^{(k)}(i,j) := \mathbb{P}(X_{n+k} = j \mid X_n = i).$$

One of the most important consequences of the Markov property is as follows.

Proposition 1 (Probability of a sequence of states I).

The one-step transition probabilities fully determine the probability of any sequence of states:

$$\mathbb{P}(X_n = x_n, \dots, X_1 = x_1 \mid X_0 = x_0) = p(x_0, x_1) \cdots p(x_{n-1}, x_n).$$

The proof is left as an exercise in the chain rule for intersections and the Markov property. Then, generalize the statement above using time-homogeneity.

Proposition 2 (Backwards Markov property).

Every Markov chain satisfies the **backwards Markov property**:

$$\mathbb{P}(X_0 = x_0 \mid X_1 = x_1, \dots, X_n = x_n) = \mathbb{P}(X_0 = x_0 \mid X_1 = x_1)$$

for every $n \in \mathbb{Z}^+$ and every feasible sequence of states $(x_0, \dots, x_n)$.

Proof. We proceed by the definition of conditional probability and Bayes' rule:

$$\begin{aligned}
\mathbb{P}(X_0 = x_0 \mid X_1 = x_1, \dots, X_n = x_n) &= \frac{\mathbb{P}(X_0 = x_0, X_1 = x_1, \dots, X_n = x_n)}{\mathbb{P}(X_1 = x_1, \dots, X_n = x_n)} \\
&\overset{*}{=} \frac{\mathbb{P}(X_0 = x_0) \cdot \mathbb{P}(X_1 = x_1 \mid X_0 = x_0) \cdot \prod_{i=1}^{n-1} p(x_i, x_{i+1})}{\mathbb{P}(X_1 = x_1) \cdot \prod_{i=1}^{n-1} p(x_i, x_{i+1})} \\
&= \mathbb{P}(X_0 = x_0 \mid X_1 = x_1).
\end{aligned}$$

□

For notational convenience, we write $X^{(m:n)}$ for $(X_m, \dots, X_n)$, where m may be greater than n. The backwards Markov property says "the immediate past X_0 and future $X^{(2:n)}$ are conditionally independent given the present," the reverse side of the coin to the Markov property.

Definition 5 (Reverse transition probabilities).

$$\tilde{p}(i, j) := \mathbb{P}(X_n = j \mid X_{n+1} = i) = \frac{\mathbb{P}(X_n = j) \cdot \mathbb{P}(X_{n+1} = i \mid X_n = j)}{\mathbb{P}(X_{n+1} = i)}.$$

Proposition 3 (Probability of a sequence of states II).

$$\mathbb{P}(X_{n-1} = x_{n-1}, \dots, X_0 = x_0 \mid X_n = x_n) = \tilde{p}(x_{n-1}, x_n) \cdots \tilde{p}(x_0, x_1).$$

Definition 6 (Reversed Markov chain).

Let N be a positive integer, and let $Y_n := X_{N-n}$, so that $(Y_0, \ldots, Y_N) := (X_N, \ldots, X_0)$. Then $(Y_n)_{n=0}^N$ is the **reversed chain** of $(X_n)_{n\in\mathbb{N}}$.

The reversed chain is indeed a Markov chain by the backwards Markov property for X_n, but Y_n is *not* time-homogeneous in general. We will delve further into chains that can "ignore" the direction of the flow of time in the section on *reversibility*.

Optionally, let us finish by formally proving a very intuitive claim.

Proposition 4 ("Fragmented memories").

The future is dependent only on the present, regardless of *any* knowledge about the past. Let $I \subseteq \{0, \ldots, n-1\}$, and define the (partial) **history at times** I to be the event

$$\left\{X^{(I)} = x^{(i)}\right\} = \{X_{i_1} = x_{i_1}, \ldots, X_{i_k} = x_{i_k}\} = \bigcap_{i\in I} \{X_i = x_i\}.$$

The *null history* or *event of no information* is $\Omega = \{X^{(\varnothing)} = x^{(\varnothing)}\}$, while the *full history* or *complete trajectory* is $\{X^{(0,\ldots,n-1)} = x^{(0,\ldots,n-1)}\}$, such that

$$\Omega \supseteq \left\{X^{(I)} = x^{(I)}\right\} \supseteq \{X_0 = x_0, \ldots, X_{n-1} = X_{n-1}\}.$$

Then, by the Markov property, the transition probability $\mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n)$ is the same whether given no history, any partial history, or the full history:

$$\begin{aligned} p(x_n, x_{n+1}) &= \mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n, \Omega) \\ &= \mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n, X^{(I)} = x^{(I)}) \\ &= \mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0). \end{aligned}$$

Proof (Alex Fu). Let $I^c = \{0, \ldots, n-1\} \setminus I$. By the law of total probability, we can take the summations over all possible values of the unknown *nuisable variables* $X^{(I^c)}$ to get

$$\begin{aligned} \mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n, X^{(I)} = x^{(I)}) &= \frac{\mathbb{P}(X_{n+1} = x_{n+1}, X_n = x_n, X^{(I)} = x^{(I)})}{\mathbb{P}(X_n = x_n, X^{(I)} = x^{(I)})} \\ &= \frac{\sum \mathbb{P}(X^{(n+1:0)} = x^{(n+1:0)})}{\sum \mathbb{P}(X^{(n:0)} = x^{(n:0)})} \\ &\overset{*}{=} \frac{\sum \left[p(x_n, x_{n+1}) \prod_{i=0}^{n-1} p(x_i, x_{i+1})\right]}{\sum \prod_{i=0}^{n-1} p(x_i, x_{i+1})} \\ &= p(x_n, x_{n+1}). \end{aligned}$$

□

3.2 Mass and flow

Definition 7 (Probability mass; probability flow).

The state space S is countable, so at every time step n, a probability distribution over S is given by the probability mass function

$$\pi_n(i) := \mathbb{P}(X_n = i), \qquad i \in S.$$

This is the (probability) **mass** of state i at time n. Then, the (probability) **flow** from state i to state j during time step n is

$$\mathbb{P}(X_n = i, X_{n+1} = j) = \pi_n(i) \cdot p(i, j).$$

In this characterization, we can almost forget any probabilities of a Markov chain, instead treating one as a network of flows with total mass 1, where $p(i,j)$ describes the proportion of mass in state i that flows to j. We can thus use intuitive vocabulary like "supporting" mass or "circulating" mass, and this analogy is strengthened by the following properties.

Proposition 5 (Conservation of mass).

The total mass in the Markov chain is always 1. For every $n \in \mathbb{N}$,

$$\sum_{i \in S} \pi_n(i) = 1.$$

Proposition 6 (Flow-in).

The total flow into a state is the sum of flows-in from every state — mass has to come from somewhere, justified by the law of total probability.

$$\pi_{n+1}(j) = \sum_{i \in S} \pi_n(i) \cdot p(i, j).$$

Proposition 7 (Flow-out).

The mass in a state always equals the total flow out (though mass can return by self-loop).

$$\pi_n(i) = \sum_{j \in S} \pi_n(i) \cdot p(i, j).$$

Equivalently, the sum of transition probabilities out of any state is 1. For any $i \in S$,

$$\sum_{j \in S} p(i, j) = 1.$$

3.3 Transition diagram

Definition 8 (Transition diagram).

Every DTMC is uniquely associated to a **transition diagram**, a directed and weighted graph possibly with self-loops, but without multiedges. The set of vertices is the state space S, and an edge (i, j) exists iff the corresponding transition probability $p(i, j)$ is nonzero.

Conversely, any graph that is directed, weighted, possibly with self-loops, without multiedges, and has edge weights in $(0, 1]$ uniquely determines a DTMC, so we often refer to the transition diagram of a DTMC as the DTMC itself.

A Markov chain is most commonly visualized as its transition diagram. A *realization* is given by a particle traversing the states $x_0, x_1, x_2, \ldots$, at each time step probabilistically choosing its next position. With a large collection of such particles, the proportion of particles in any given state approaches the probability *mass* of the state.

Moreover, the use of transition diagrams allows us to harness the language of *graph theory*.

Definition 9 (Terminology of transition diagrams).

a. State j is **reachable** or *accessible* from state i, denoted $i \to j$, if there exists a path from i to j in the transition diagram.

b. States i and j **communicate**, denoted $i \leftrightarrow j$, if they are each reachable from the other. That is, the vertices i and j are *strongly connected*.

c. The **communicating class** or *strongly connected component* (SCC) $[i]$ of state i is the collection of all states communicating with i. A communicating class also forms a *subgraph* of the transition diagram.

d. The relation of communicating $\leftrightarrow$ is also an *equivalence relation* on the state space. The communicating classes *partition* the state space: every state belongs to one and only one communicating class, which are disjoint from each other.

e. A Markov chain is **irreducible** if its transition diagram has only one communicating class (itself) or equivalently is strongly connected. Otherwise, the chain is **reducible**. Note that every communicating class is itself irreducible (as a subgraph).

f. *Every directed graph is a directed acyclic graph (DAG) of SCCs*, so every transition diagram is also a DAG of communicating classes. A class is a **source** if no edges lead into it; a **sink** or **closed** if no edges lead out it; and **isolated** if both are true. A *state* is **absorbing** if no edges lead out of it.

g. A **class property** is any property that holds the same for every state in a communicating class — if one state has it, then every state in the class has it. Central examples include *positive recurrence*, *null recurrence*, *transience*, *aperiodicity*, and *ergodicity*.

A general principle: we will mostly work with class properties in irreducible Markov chains, as any Markov chain can be decomposed into (irreducible) classes, and the properties of each class can usually be combined in a straightforward manner.

3.4 Linear stochastic system

Definition 10 (Stochastic vector; row-stochastic matrix).

A **stochastic vector** is a row vector π with $|S|$ nonnegative real entries which sum to 1:

$$\sum_{i=1}^{|S|} \pi(i) = 1.$$

A (right- or row-) **stochastic matrix** is a square matrix P with nonnegative real entries such that every row is a stochastic vector, or every row sums to 1:

$$\sum_{j=1}^{|S|} P_{i,j} = 1, \qquad i = 1, \ldots, |S|\,.$$

Then a DTMC is uniquely determined by a *stochastic matrix-vector equation* with probability distribution vector π_n and transition probability matrix P:

$$\pi_{n+1} = \pi_n P.$$

The ith entry of the row vector π_n is equal to the value of the distribution π_n at state i, and the (i,j)th entry in the matrix P is equal to the transition probability $p(i,j)$. Thus the notation $\pi_n(i) = \pi_n(i)$ and $P_{i,j} = p(i,j)$ is unambiguous.

Where does the linearity in this "linear" system come from? If we consider a specific entry in the vector-matrix multiplication,

$$\pi_{n+1}(j) = \pi_n \cdot \mathsf{row}_j(P) = \sum_{i \in S} \pi_n(i) \cdot p(i,j),$$

which is precisely the law of total probability or the flow-in equation.

Proposition 8 (Convex combinations of stochastic vectors).

Stochastic vectors are not closed under general linear combinations, because the resulting entries may not sum to 1. However, if $v_1, \ldots, v_n$ are stochastic vectors, and $\alpha_1, \ldots, \alpha_n$ are nonnegative coefficients that sum to 1, then the **convex combination**

$$\alpha_1 v_1 + \ldots + \alpha_n v_n$$

remains a stochastic vector.

Convex combinations will be key in combining distributions over individual communicating classes.

By induction, $\pi_n = \pi_0 P^n$ for every $n \in \mathbb{N}$. Thus a Markov chain as a simple random process is completely specified by $|S|^2 - 1$ values in $[0,1]$: the initial distribution requires $|S| - 1$, and the transition probabilities require $|S|^2 - |S|$.

Proposition 9 (k-step transition probabilities).

Another nice consequence of the row-stochastic convention is that

$$p^{(k)}(i,j) = \mathbb{P}(X_k = j \mid X_0 = i) = (P^k)_{i,j}.$$

By the law of total probability, summing over all possible values of the *nuisance variables* $X^{(1:k-1)}$, or all possible immediate sequences of length k,

$$\begin{aligned} p^{(k)}(i,j) &= \sum_{x^{(1:k-1)}} \mathbb{P}\big(X_k = j, X^{(1:k-1)} = x^{(1:k-1)} \mid X_0 = i\big) \\ &= \sum_{x_1,\dots,x_{k-1}} p(x_{k-1}, j) \cdot p(x_{k-2}, x_{k-1}) \cdots p(x_1, x_2) \cdot p(i, x_1) \\ &= (P^k)_{i,j}. \end{aligned}$$

Proposition 10 (Chapman-Kolmogorov equations).

For all $k, \ell \in \mathbb{N}$, $P^{k+\ell} = P^k P^\ell$.

From our setup, this may seem like a simple identity perhaps not worthy of being specially named, but its underlying importance is that we can compute any $(k+\ell)$-step transition probability given just the k-step and ℓ-step transition probabilities.

We finish with one of many interesting out-of-scope connections between the graph theory and linear algebra perspectives of Markov chains.

Definition 11 (Irreducible matrix*).

A matrix A is **irreducible** if there does not exist a permutation matrix Π such that A is similar to a block upper triangular matrix under conjugation by Π, and **reducible** otherwise.

Suppose that the transition probability matrix P is reducible, with the permutation matrix given by the identity I. Then $IPI^{-1} = P$ is block upper triangular:

$$P = \begin{bmatrix} [*] & & * \\ & \ddots & \\ 0 & & [*] \end{bmatrix}.$$

So there exists a partition of the state space into “blocks,” with zero probability of reaching one block from another, even if the probability in the other direction is nonzero — the transition diagram is reducible! Then, the permutation matrix Π simply permutes the order of states in P without changing any transition probabilities between states, so we have

Proposition 11 (Equivalence of irreducibilities).

A Markov chain is irreducible if and only if its transition probability matrix P is irreducible.

Continued in part II.

■