1 First look: Markov chains as stochastic systems

1.1 Stochastic signals (a.k.a. stochastic processes)

A deterministic signal is just a real- or a complex-valued function of time. Let us focus on real-valued signals, to keep things concrete. In that case, a signal x is just a function $x : T \to \mathbb{R}$, where the points in its domain T takes integer valueds (in which case we have a *discrete-time* signal) or real values (in which case case we have a *continuous-time* signal). The notation $x : T \to \mathbb{R}$ is meant to distinguish the *entire* signal, which is a *function*, from its *value* $x(t)$ at a given time t. Another way of viewing the signal x is as a collection of the values it takes, indexed by the corresponding times: we write this as $x = (x(t))_{t\in T}$. Since the signal is deterministic, its value $x(t)$ at each $t \in T$ is well-defined and can be computed exactly (at least in principle).

A *stochastic signal*, by contrast, is a collection of *random variables*, so we write it as $X = (X_t)_{t\in T}$. Note two notational differences: we are using uppercase letters for random quantities and lowercase ones for deterministic quantities, and we are also writing the time t as a subscript, instead of listing it in parentheses. Now, every time we encounter random variables, there is a *probability space*[1] lurking behind them. Recall that a probability space is a triple $(\Omega, \mathcal{F}, \mathbf{P})$, where Ω is the sample space, $\mathcal{F}$ is a collection of events, i.e., certain distinguished subsets of Ω, and $\mathbf{P}$ is the probability measure that assigns a number between 0 and 1 to each event $E \in \mathcal{F}$. A single random variable is a function $X : \Omega \to \mathbb{R}$ — it associates a real value $X(\omega)$ to each element ω of the sample space Ω. Now, a stochastic signal X is a collection $(X_t)_{t\in T}$ of random variables indexed by the elements t of T. So, we have a random variable $X_t : \Omega \to \mathbb{R}$ for each $t \in T$, and all of these random variables taken together constitute the description of the stochastic signal X. In probability theory, such collections of random variables defined on a common probability space are called *stochastic processes*. From our engineering perspective, though, it is more convenient to think of them as signals.

This formulation is actually rather intuitive once we remember that a stochastic signal is a signal that cannot be specified exactly because of some chance effects. The influence of these chance effects comes through the probability space $(\Omega, \mathcal{F}, \mathbf{P})$ — to an ideal observer who knows the value of ω, there is nothing random about X, since the value of the signal at each time t is given by $X_t(\omega)$. Unlike the ideal observer, we are in a situation when we do not know ω, but can only assign probabilities to various events associated with it. So, for example, we can compute the probability that the value X_t falls between a and b,

$$\mathbf{P}(a \le X_t \le b) = \mathbf{P}\left(\{\omega \in \Omega : a \le X_t(\omega) \le b\}\right),$$

or, more generally, probabilities of the form

$$\begin{aligned}\mathbf{P}(a_1 \le X_{t_1} \le b_1, \ldots, a_n \le X_{t_n} \le b_n) &= \mathbf{P}\left(\left\{\omega \in \Omega : a_i \le X_{t_i}(\omega) \le b_i,\, 1 \le i \le n\right\}\right) \\ &= \mathbf{P}\left(\bigcap_{i=1}^{n} \left\{\omega \in \Omega : a_i \le X_{t_i}(\omega) \le b_i\right\}\right)\end{aligned}$$

for a collection of times $t_1, \ldots, t_n \in T$. We will see later that this will provide a complete probabilistic description of the stochastic signal X.

[1] Now is as good a time as any to dust off your ECE 313 lecture notes and refresh your memory about the axioms of probability and the definition of a random variable.

1.2 Deterministic systems with a state

We are interested in the *dynamics* of signals, i.e., in the way they unfold or evolve in time. Let's look at a deterministic signal $x : T \to \mathbb{R}$. To keep things simple, we assume that x is a discrete-time signal, and take $T = \mathbb{Z}_+ \triangleq \{0, 1, 2, \ldots\}$. The most general dynamical description would allow us to compute the value $x(t+1)$ for an arbitary $t \in T$ in terms of the *history* $(x(0), \ldots, x(t))$:

$$x(t+1) = f_t(x(0), \ldots, x(t)).$$

Here, f_t is the *update rule* at time t, and the superscript t is meant to suggest that the update rule will generally depend on t. Such a description is very cumbersome: in order to determine the value of the signal at time t, we need to keep track of the entire history of the signal. Even in the simplest case of a binary signal, i.e., when $x(t) \in \{0, 1\}$ for each t, there are 2^t possible histories at time t! For this reason, we prefer dynamical descriptions that are more compact. For example, suppose that the value of the signal $x(t+1)$ depends only on $x(t)$, but not on the rest of the history:

$$x(t+1) = f_t(x(t)), \qquad t = 0, 1, \ldots. \tag{1.1}$$

This drastically simplifies the task of tracking the dynamics: instead of keeping the entire history, we only keep track of the most recent value $x(t)$, and we also need a clock to know what time it is. If the update rule does not depend on t, i.e., if there exists some function $f : \mathbb{R} \to \mathbb{R}$ such that

$$x(t+1) = f(x(t)), \qquad t = 0, 1, \ldots \tag{1.2}$$

then the system is *time-invariant*, and we only need to store the most recent value of x. When we are in a situation described by (1.1) or (1.2), the update functions $(f_t)_{t \in T}$ (in the time-varying case) or the single update function f (in the time-invariant face) define a *dynamical system with a state*, and we say that $x(t)$ is the state of the system at time t. In general, we use the word "state" to refer to any variable or collection of variables that summarize all relevant information at time t. In our case, relevant information pertains to predicting the value $x(t+1)$ of the signal at time t, and the most recent value $x(t)$ is the state variable — the rest of the history, i.e., $x(0), \ldots, x(t-1)$, is irrelevant.

The continuous-time counterparts of such systems with a state are described by first-order ordinary differential equations (ODEs). These take the form

$$\frac{\mathrm{d}}{\mathrm{d}t} x(t) = f(x(t), t), \qquad t \geq 0 \tag{1.3}$$

with a given initial condition $x(0)$, or, in the time-invariant case,

$$\frac{\mathrm{d}}{\mathrm{d}t} x(t) = f(x(t)). \tag{1.4}$$

To see why (1.3) and (1.4) are continuous-time counterparts of (1.1) and (1.2), suppose that we sample the value of $x(t)$ every δ seconds, starting at $t = 0$. Thus, instead of the full trajectory $x(t)$,

$t \geq 0$, we observe the samples $\tilde{x}(k) \triangleq x(k\delta)$, for $k = 0, 1, 2, \ldots$. This is a discrete-time signal, and, if δ is small enough, we can write down the following discrete-time approximations of (1.3) and (1.4):

$$\tilde{x}(k+1) \approx \tilde{x}(k) + \delta \cdot \tilde{f}_k(\tilde{x}(k)),$$

and

$$\tilde{x}(k+1) \approx \tilde{x}(k) + \delta \cdot \tilde{f}(\tilde{x}(k)),$$

where we have defined $\tilde{f}_k(\tilde{x}(k)) \triangleq f(x(k\delta), k\delta)$ and $\tilde{f}(\tilde{x}(k)) \triangleq f(x(k\delta))$. These equations show that the evolution of the sampled trajectory $(\tilde{x}(k))_{k\in\mathbb{Z}_+}$ can be approximated by a dynamical model of the form (1.1) or (1.2). In many cases, the initial description of the system is given by a higher-order ODE. In such a case, it is convenient to expand the state space and write down several 1st-order ODEs. For example, in mechanics we may want to describe the motion of a particle subject to a force. Let $x(t)$ denote the position of the particle of mass m at time t, and let $F = F(x(t), t)$ be a position- and time-dependent force experienced by the particle. Then Newton's law tells us that

$$\ddot{x}(t) = \frac{F(x(t), t)}{m}, \tag{1.5}$$

where $\dot{x}$ denotes the derivative with respect to t, and $\ddot{x}$ denotes the second derivative with respect to t. Suppose we track the position of the particle every δ seconds starting from $t = 0$, where $\delta > 0$ is some small number – i.e, we are interested in $\tilde{x}(k)$ for $k \in \mathbb{Z}_+$. In that case, we can write down the following discrete-time approximation of (1.5):

$$\frac{\tilde{x}(k+2) - 2\tilde{x}(k+1) + \tilde{x}(k)}{\delta^2} \approx \frac{\tilde{F}_k(\tilde{x}(k)}{m},$$

with the appropriately defined $\tilde{F}_k$. This shows that, in order to predict the position of the particle at time $t = k\delta$, we need to remember its position at times $(k-1)\delta$ and $(k-2)\delta$. However, by augmenting our description of the particle to include its *velocity* at each time $k\delta$, we can get away with only remembering the most recent information: Recalling that $v(t) = \dot{x}(t)$, we can write two first-order ODEs:

$$\begin{aligned} \dot{x}(t) &= v(t) \\ \dot{v}(t) &= \frac{F(x(t), t)}{m}. \end{aligned}$$

Now, discretizing the time in units of δ, we can write

$$\begin{aligned} \tilde{x}(k+1) &\approx \tilde{x}(k) + \delta\tilde{v}(k) \\ \tilde{v}(k+1) &\approx \tilde{v}(k) + \frac{\tilde{F}_k(\tilde{x}(k))}{m}. \end{aligned}$$

In this way, the position and the velocity values at each time t constitute the state of the particle.

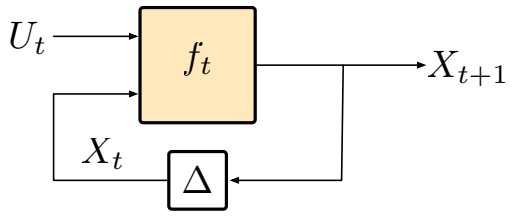

Figure 1: Markov chain as a stochastic system with a state; Δ is the unit delay: $\Delta(X_{t+1}) = X_t$.

1.3 Markov chains: stochastic systems with a state

We will now extend our definition of the system with a state to the stochastic case. The simplest nontrivial way of doing that is to inject some randomness into the state update rule. For now, we will work in discrete time and take $T = \mathbb{Z}_+$. Let $(U_t)_{t\in\mathbb{Z}_+}$ be a sequence of independent and identically distributed (i.i.d.) random variables, and consider a stochastic signal $X = (X_t)_{t\in\mathbb{Z}_+}$ that evolves according to the following rule:

$$X_{t+1} = f_t(X_t, U_t), \qquad t = 0, 1, 2, \ldots. \tag{1.6}$$

Here, f_t is the update rule at time t, and the initial condition X_0 may be deterministic or stochastic, but we assume that it is also independent of $U_0, U_1, \ldots$. Any stochastic signal X that evolves according to (1.6) is called a *Markov chain*. When the update rule does not depend on t, we have a simpler model

$$X_{t+1} = f(X_t, U_t), \qquad t = 0, 1, 2, \ldots \tag{1.7}$$

and we say that X is a *time-homogeneous Markov chain*. Otherwise, X is *time-inhomogeneous*. Figure 1 shows a block diagram corresponding to (1.6).

Important! Independence of the random variables $U_0, U_1, \ldots$ is crucial.

For the rest of the lecture, we will keep things simple and focus on time-homogeneous Markov chains. Moreover, we will assume that the random variables X_t take integer values, in which case we are dealing with a *discrete-state* Markov chain. We will denote the set of all possible values X_t can take at any time t by X, and we will call it the *state space* of the Markov chain.

At this point, we see that the main difference between the deterministic system (1.1) and its stochastic counterpart (1.6) is the presence of the random input U_t. That is, in order to produce the next value X_{t+1}, we generate a fresh random input U_t and apply the update rule f_t to the current value X_t and to U_t. It somehow feels right to call X_t the *state* of the Markov chain, but let's make this intuition precise. To do that, let us compute the probability $\mathbf{P}[X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n]$ that X goes through a given sequence of values $x_0, x_1, \ldots, x_n \in \mathsf{X}$ at times $t = 0, 1, \ldots, n$. Because of the update rule (1.7), we can write this probability as

$$\mathbf{P}[X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n] = \mathbf{P}[X_0 = x_0, f(x_0, U_0) = x_1, \ldots, f(x_{n-1}, U_{n-1}) = x_n].$$

Note that there are two sources of randomness here: the initial condition X_0 and the stochastic inputs $U_0, \ldots, U_{n-1}$. These random variables are independent, so we can further write

$$\mathbf{P}[X_0 = x_0, f(x_0, U_0) = x_1, \ldots, f(x_{n-1}, U_{n-1}) = x_n] = \mathbf{P}[X_0 = x_0] \prod_{t=0}^{n-1} \mathbf{P}[f(x_t, U_t) = x_{t+1}]. \tag{1.8}$$

Now, assuming that $x_0, \ldots, x_n$ are such that $\mathbf{P}[X_t = x_t, 0 \le t \le n]$ is nonzero, we can ask for the *conditional probability* that X_{n+1} takes some fixed value x_{n+1}, given $X_0 = x_0, \ldots, X_n = x_n$. By definition of conditional probabilities, we have

$$\mathbf{P}[X_{n+1} = x_{n+1} | X_t = x_t, 0 \le t \le n] = \frac{\mathbf{P}[X_0 = x_0, \ldots, X_n = x_n, X_{n+1} = x_{n+1}]}{\mathbf{P}[X_0 = x_0, \ldots, X_n = x_n]}.$$

The denominator is given by (1.8), and we also already know how to compute the numerator — use (1.8), but with time going up to $n+1$:

$$\begin{aligned}
&\frac{\mathbf{P}[X_0 = x_0, \ldots, X_n = x_n, X_{n+1} = x_{n+1}]}{\mathbf{P}[X_0 = x_0, \ldots, X_n = x_n]} \\
&\quad = \frac{\mathbf{P}[X_0 = x_0] \prod_{t=0}^{n} \mathbf{P}[f(x_t, U_t) = x_{t+1}]}{\mathbf{P}[X_0 = x_0] \prod_{t=0}^{n-1} \mathbf{P}[f(x_t, U_t) = x_{t+1}]} \\
&\quad = \frac{\mathbf{P}[X_0 = x_0] \prod_{t=0}^{n-1} \mathbf{P}[f(x_t, U_t) = x_{t+1}] \mathbf{P}[f(x_n, U_n) = x_{n+1}]}{\mathbf{P}[X_0 = x_0] \prod_{t=0}^{n-1} \mathbf{P}[f(x_t, U_t) = x_{t+1}]} \\
&\quad = \mathbf{P}[f(x_n, U_n) = x_{n+1}].
\end{aligned}$$

What do we see? We see that the conditional probability distribution of X_{n+1} given the entire history $X_0 = x_0, \ldots, X_n = x_n$ depends only on the most recent value $X_n = x_n$, and not on anything else. This is what gives us the license to refer to X_t as the *state* of the Markov chain. Similar reasoning also shows that, for any pair of possible states $x, y \in \mathsf{X}$, we have

$$\mathbf{P}[X_{t+1} = y | X_t = x] = \mathbf{P}[f(x, U_t) = y].$$

Now, let us examine the quantity $\mathbf{P}[f(x, U_t) = y]$ more closely. Since our Markov chain is time-homogeneous, we can write

$$\mathbf{P}[f(x, U_t) = y] = \mathbf{P}[f(x, U_0) = y],$$

Now, if we define the shorthand notation $M(x, y) \triangleq \mathbf{P}[f(x, U) = y]$, we can rewrite (1.8) as

$$\mathbf{P}[X_0 = x_0, \ldots, X_n = x_n] = \mathbf{P}[X_0 = x_0] \cdot \prod_{t=0}^{n-1} M(x_t, x_{t+1}). \tag{1.9}$$

What can we say about the quantities $M(x, y)$? Being defined in terms of probabilities, they are bounded between 0 and 1. Moreover, for any x we have

$$\sum_y M(x, y) = 1. \tag{1.10}$$

To justify this, recall the probabilistic interpretation $M(x,y) = \mathbf{P}[X_{t+1} = y | X_t = x]$, i.e., $M(x,y)$ is the conditional probability that the state at time $t+1$ is y, given that the state at time t was x. Since the chain will reach *some* state starting at x, we must have (1.10). The parameters $M(x,y)$ are called the *state transition probabilities* of the Markov chain, and they are determined by the state update rule f and by the common distribution of $U_0, U_1, \ldots$, and we will see concrete examples of this later on.

Now, the way we have defined a Markov chain is through the dynamical model (1.7). A moment of reflection shows that this gives an *imperative* description of the Markov chain: if we have a subroutine for generating independent random variables $U_0, U_1, \ldots$ with a prescribed distribution, then we can write computer code to generate the state trajectory $X_0, X_1, \ldots$ of the Markov chain. On the other hand, specifying the transition probabilities $M(x,y)$ is a *declarative* description: it tells us where the chain is likely to be at any given time t, but it does not tell us *how* it can get there. Now let us look at a couple of concrete examples.

1.4 A two-state Markov chain

A classic model of a Markov chain is a chain with two states, thus $\mathsf{X} = \{0,1\}$. We will look at the following example: let $U_0, U_1, \ldots$ be uniformly distributed on the unit interval $[0,1]$, and consider the state update rule

$$f(0,u) = \begin{cases} 0, & 0 \le u < 2/3 \\ 1, & 2/3 \le u < 1 \end{cases}, \qquad f(1,u) = \begin{cases} 1, & 0 \le u < 2/3 \\ 0, & 2/3 \le u < 1 \end{cases}. \tag{1.11}$$

This state transition rule is pictured in Fig. 2. From this imperative description, we can compute

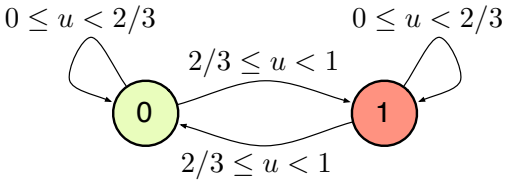

Figure 2: The transition diagram for a two-state Markov chain.

the declarative description in terms of state transition probabilities. For example,

$$M(0,1) = \mathbf{P}[X_1 = 1 | X_0 = 0] = \mathbf{P}[2/3 \le U_0 < 1] = 1/3, \tag{1.12}$$

where we have used the fact that each U_t is a Uniform$(0,1)$ random variable, which means that, for any $0 \le a \le b < 1$, $\mathbf{P}[a \le U < b] = b - a$. We can fill in the rest of the transition probabilities in the same way, but we can also exploit the obvious symmetry in the transition diagram in Fig. 2 and write

$$M(x, x \oplus 1) = 1 - M(x,x) = \frac{1}{3}, \qquad \forall x \in \{0,1\}. \tag{1.13}$$

In other words, the next state X_{t+1} is equal to the current state X_t with probability $\frac{2}{3}$, and flips with probability $\frac{1}{3}$. This is an example of a *finite-state* Markov chain.

The computation of probabilities amounts to multiplying vectors by matrices. To see this, for each $t \in \mathbb{Z}_+$ let us define a row vector $p_t = (p_t(0), p_t(1))$ corresponding to the probability distribution of the state X_t at time t. Then, for any $x \in \{0,1\}$ we can write

$$\begin{aligned} p_{t+1}(x) &= \sum_{x' \in \{0,1\}} \mathbf{P}[X_t = x', X_{t+1} = x] \\ &= \sum_{x' \in \{0,1\}} \mathbf{P}[X_t = x'] \mathbf{P}[X_{t+1} = x | X_t = x'] \\ &= \sum_{x' \in \{0,1\}} p_t(x') M(x', x). \end{aligned}$$

Using (1.13), we can express this in a compact matrix form:

$$\begin{aligned} \begin{pmatrix} p_{t+1}(0) & p_{t+1}(1) \end{pmatrix} &= \begin{pmatrix} p_t(0) & p_t(1) \end{pmatrix} \begin{pmatrix} M(0,0) & M(0,1) \\ M(1,0) & M(1,1) \end{pmatrix} \\ &= \begin{pmatrix} p_t(0) & p_t(1) \end{pmatrix} \begin{pmatrix} 2/3 & 1/3 \\ 1/3 & 2/3 \end{pmatrix}. \end{aligned}$$

Denoting the 2×2 matrix of the state transition probabilities by M, we obtain the one-step update equation

$$p_{t+1} = p_t M, \qquad t = 0, 1, 2, \ldots. \tag{1.14}$$

Iterating it, we can compute the state probability distribution p_t at any time t from the initial distribution p_0:

$$p_t = p_0 M^t, \tag{1.15}$$

where M^t is the tth power of the matrix M.

1.5 Random walk on the integers

Let us look at an example of a Markov chain with a countably infinite state space. Consider a particle that moves in unit steps on the real line as follows: if at time $t = 0, 1, 2, \ldots$ the particle is at location $x \in \mathbb{Z}$, then at time $t+1$ it hops one step to the left (i.e., to $y = x - 1$) or one step to the right (i.e., to $y = x + 1$) with equal probability. We can describe the motion of the particle as follows. Let $U_0, U_1, \ldots$ be i.i.d. random variables, where each U_t takes values ± 1 with equal probability. (Such random variables are often referred to as *Rademacher random variables*.) Then the integer-valued state X_t evolves according to the rule

$$X_{t+1} = X_t + U_t, \qquad t = 0, 1, 2, \ldots.$$

So, according to our definitions, this is a time-homogeneous Markov chain with state space $\mathsf{X} = \mathbb{Z}$ and with the update rule $f(x, u) = x + u$. This Markov chain is often referred to as a *simple symmetric random walk* on the integers.

We can immediately write down the state transition probabilities:

$$M(x,y) = \mathbf{P}[x + U_0 = y] = \mathbf{P}[U_0 = y - x] = \begin{cases} \frac{1}{2}, & y - x = \pm 1 \\ 0, & \text{otherwise} \end{cases}.$$

Using this expression, we can use (1.9) to compute various probabilities associated with the trajectory of the particle. For example, if we denote by p_t the pmf of X_t, the state of the random walk at time t, then $p_{t+1} = p_t M$, or, for any $y \in \mathsf{X}$,

$$\begin{aligned} p_{t+1}(y) &= \sum_{x \in \mathbb{Z}} p_t(x) M(x,y) \\ &= \frac{1}{2} p_t(y-1) + \frac{1}{2} p_t(y+1). \end{aligned} \tag{1.16}$$

Similarly,

$$\begin{aligned} p_{t+2}(y) &= \sum_{x \in \mathbb{Z}} p_{t+1}(x) M(x,y) \\ &= \frac{1}{2} p_{t+1}(y-1) + \frac{1}{2} p_{t+1}(y+1) \\ &= \frac{1}{2}\left(\frac{1}{2} p_t(y-2) + p_t(y)\right) + \frac{1}{2}\left(\frac{1}{2} p_t(y) + p_t(y+2)\right) \\ &= \frac{1}{4} p_t(y-2) + \frac{1}{2} p_t(y) + \frac{1}{2} p_t(y+2). \end{aligned} \tag{1.17}$$

We can also analyze the behavior of the random walk by looking at expected values of various functions of the state. That is, consider some function $g : \mathbb{Z} \to \mathbb{R}$, which is *absolutely summable*, i.e.,

$$\sum_{x \in \mathbb{Z}} |g(x)| < \infty. \tag{1.18}$$

Just like we were representing all pmf's on X as row vectors, we can represent all such functions as *column vectors*, and then

$$\begin{aligned} \mathbf{E}[g(X_0)] &= \sum_{x \in \mathbb{Z}} p_0(x) g(x) \\ &= p_0 g. \end{aligned}$$

If (1.18) holds, then

$$\begin{aligned} |\mathbf{E}[g(X_0)]| &= \left| \sum_{x \in \mathbb{Z}} p_0(x) g(x) \right| \\ &\le \sum_{x \in \mathbb{Z}} p_0(x) |g(x)| \\ &\le \sum_{x \in \mathbb{Z}} |g(x)| \\ &< \infty, \end{aligned}$$

where we have used the fact that $0 \leq p_0(x) \leq 1$ for all x. Moreover, using (1.16), we can write

$$\begin{aligned}\mathbf{E}[g(X_1)] &= \sum_{y\in\mathbb{Z}} p_1(y)g(y) \\ &= \sum_{y\in\mathbb{Z}}\left(\sum_{x\in\mathbb{Z}} p_0(x)M(x,y)\right)g(y) \\ &= \sum_{x\in\mathbb{Z}} p_0(x)\left(\sum_{y\in\mathbb{Z}} M(x,y)g(y)\right),\end{aligned}$$

and, again, $|\mathbf{E}[g(X_1)]| < \infty$ using the same argument as the one to show that $|\mathbf{E}[g(X_0)]|$ was finite. Now, the quantity in parentheses is simply $Mg(x)$, the entry of the column vector Mg in position x. Thus, we can write

$$\begin{aligned}\mathbf{E}[g(X_1)] &= \sum_{x\in\mathbb{Z}} p_0(x)Mg(x) \\ &= p_0Mg.\end{aligned}$$

Continuing in this fashion, we derive a nice formula for the expected value $\mathbf{E}[g(X_t)]$:

$$\mathbf{E}[g(X_t)] = p_0M^tg. \tag{1.19}$$

In other words, if we want to compute the expected value of $g(X_t)$, for any t, we only need to know the initial distribution p_0 and the transition probability matrix M. Moreover, for any function $g : \mathbb{Z} \to \mathbb{R}$, Mg is another such function. This function has a meaning: recalling that $M(x,y) = \mathbf{P}[X_{t+1} = y|X_t = x]$, we see that

$$\begin{aligned}Mg(x) &= \sum_{y\in\mathbb{Z}} M(x,y)g(y) \\ &= \sum_{y\in\mathbb{Z}} g(y)\mathbf{P}[X_{t+1} = y|X_t = x] \\ &= \mathbf{E}[g(X_{t+1})|X_t = x],\end{aligned}$$

i.e., the expected value of g with respect to the conditional pmf of X_{t+1} given $X_t = x$. The computa-

tion of $M^t g$ proceeds along the same lines as in Eqs. (1.16) and (1.17). For instance,

$$\begin{aligned} Mg(x) &= \sum_{y\in\mathbb{Z}} M(x,y)g(y) \\ &= \frac{1}{2}g(x-1) + \frac{1}{2}g(x+1), \\ M^2 g(x) &= M(Mg)(x) \\ &= \sum_{y\in\mathbb{Z}} M(x,y)Mg(y) \\ &= \frac{1}{2}Mg(x-1) + \frac{1}{2}Mg(x+1) \\ &= \frac{1}{4}g(x-2) + \frac{1}{2}g(x) + \frac{1}{4}g(x+2), \end{aligned}$$

and so on. In principle, we can keep going, but the formulas quickly get out of hand. However, as we will now see, there is a cleaner way of analyzing the behavior of this Markov chain using the *discrete-time Fourier transform* (DTFT).

Recall that the DTFT of a real-valued function $g : \mathbb{Z} \to \mathbb{R}$ is defined as

$$\widehat{g}(\omega) := \sum_{x\in\mathbb{Z}} g(x)e^{-\mathrm{i}\omega x}, \tag{1.20}$$

where the frequency ω takes values in $[-\pi, \pi]$. In order for (1.20) to be well-defined, it suffices for g to be absolutely summable, as in (1.18). Indeed, if g satisfies (1.18), then

$$\begin{aligned} |\widehat{g}(w)| &\le \sum_{x\in\mathbb{Z}} |g(x)e^{-\mathrm{i}\omega x}| \\ &= \sum_{x\in\mathbb{Z}} |g(x)| \\ &< \infty \end{aligned}$$

for all $\omega \in [-\pi, \pi]$. To recover g from its DTFT $\widehat{g}$, we use the inversion formula

$$g(x) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \widehat{g}(\omega)e^{\mathrm{i}\omega x}\mathrm{d}\omega. \tag{1.21}$$

Let us compute the DTFT of Mg, for an absolutely summable function $g : \mathbb{Z} \to \mathbb{R}$:

$$\begin{aligned}
\widehat{Mg}(\omega) &= \sum_{x\in\mathbb{Z}} Mg(x)e^{-\mathrm{i}\omega x} \\
&= \sum_{x\in\mathbb{Z}} \left(\frac{1}{2}g(x-1) + \frac{1}{2}g(x+1)\right)e^{-\mathrm{i}\omega x} \\
&= \frac{1}{2}\sum_{x\in\mathbb{Z}} g(x-1)e^{-\mathrm{i}\omega x} + \frac{1}{2}\sum_{x\in\mathbb{Z}} g(x+1)e^{-\mathrm{i}\omega x} \\
&= \frac{1}{2}e^{-\mathrm{i}\omega}\sum_{x\in\mathbb{Z}} g(x)e^{-\mathrm{i}\omega x} + \frac{1}{2}e^{\mathrm{i}\omega}\sum_{x\in\mathbb{Z}} g(x)e^{-\mathrm{i}\omega x} \\
&= \frac{e^{-\mathrm{i}\omega} + e^{\mathrm{i}\omega}}{2}\widehat{g}(\omega) \\
&= \cos\omega \cdot \widehat{g}(\omega).
\end{aligned}$$

Thus, the effect of multiplying g on the left by M amounts to multipliyng $\widehat{g}(\omega)$ by $\cos\omega$. Similarly,

$$\begin{aligned}
\widehat{M^2 g}(\omega) &= \widehat{M(Mg)}(\omega) \\
&= \cos\omega \cdot \widehat{Mg}(\omega) \\
&= (\cos\omega)^2 \cdot \widehat{g}(\omega).
\end{aligned}$$

Therefore, for any $t \in \mathbb{Z}_+$,

$$\widehat{M^t g}(\omega) = (\cos\omega)^t \widehat{g}(\omega). \tag{1.22}$$

Using (1.22) and the DTFT inversion formula (1.21), we arrive at the following remarkable formula: for any absolutely summable function $g : \mathbb{Z} \to \mathbb{R}$, for any $t \in \mathbb{Z}_+$, and any $x \in \mathbb{Z}$,

$$\mathbf{E}[g(X_t)|X_0 = x] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \widehat{g}(\omega)(\cos\omega)^t e^{\mathrm{i}\omega x}\mathrm{d}\omega. \tag{1.23}$$

As an example, take

$$g(x) = \delta_a(x) := \begin{cases} 1, & \text{if } x = a \\ 0, & \text{otherwise} \end{cases}$$

for some fixed $a \in \mathbb{Z}$. Then

$$\begin{aligned}
\mathbb{E}[g(X_t)|X_0 = x] &= \sum_{y\in\mathbb{Z}} \mathbf{P}[X_t = y|X_0 = x]\delta_a(y) \\
&= \mathbf{P}[X_t = a|X_0 = x].
\end{aligned}$$

If $t = 1$, then this conditional probability is nonzero if and only if $a = x \pm 1$; if $t = 2$, it is nonzero if and only if $a \in \{x-2, x, x+2\}$, etc. But, for large t, there are many different ways of starting at

$X_0 = x$ and ending up at some other point a at time t via some sequence of left and right moves. However, we can use (1.20) to write down an explicit expression for any t and any a. First, it is easy to see that $\widehat{\delta_a}(\omega) = e^{-\mathrm{i}\omega a}$. Thus, using (1.23), we can write

$$\mathbf{P}[X_t = a | X_0 = x] = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{\mathrm{i}\omega(x-a)} (\cos\omega)^t \, \mathrm{d}\omega. \tag{1.24}$$

In particular, if $x = a$, then we are asking for the probability of returning to the starting point at time t: for any $a \in \mathbb{Z}$,

$$\mathbf{P}[X_t = a | X_0 = a] = \frac{1}{2\pi} \int_{-\pi}^{\pi} (\cos\omega)^t \, \mathrm{d}\omega.$$

In this particular case, though, we can compute this conditional probability directly. First note that there is no way for the particle to return to its starting point in an odd number of steps — indeed, if $X_0 = a$, then $X_t = a \pm 1$ at any odd t. Thus, $\mathbf{P}[X_t = a | X_0 = a] = 0$ if t is odd. On the other hand, if t is even, then any of moves that has the same number of left and right moves will result in the particle returning to its position at time 0. Now, at each time step the particle moves one step to the left or one step to the right with equal probability, and there are $\binom{t}{t/2}$ distinct patterns of t moves that have an equal number of left and right moves. Therefore, $\mathbf{P}[X_t = a | X_0 = a] = \frac{1}{2^t}\binom{t}{t/2}$ for t even. Thus, we have derived a nice identity:

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} (\cos\omega)^t \, \mathrm{d}\omega = \begin{cases} 0, & \text{if } t \text{ is odd} \\ \frac{1}{2^t}\binom{t}{t/2}, & \text{if } t \text{ is even} \end{cases}.$$

1.6 Long-term behavior of a Markov chain

As we have just seen, we can arrange the one-step transition probabilities $M(x, y)$ of any Markov chain into a matrix whose rows and columns are indexed by the elements of the state space X, and whose entry in row x, column y is given by $M(x, y)$. Now the calculation of the state distribution p_t at any time t from a given initial distribution p_0 at time 0 is just matrix multiplication, as we have seen in (1.15). So, it is natural to ask: What happens when t grows very large? In light of (1.15), it is clear that the answer has to do with the behavior of the powers of M, and in particular with what happens to M^t as $t \to \infty$.

Now, the theory of long-term behavior of Markov chains is surprisingly rich and complex, and we will look into some of it later on. For now, we will limit ourselves to simple heuristic arguments. Let us take a look at the two-state Markov chain from Section 1.4. We can compute the first few

powers of M, just to see what happens:

$$M = \begin{pmatrix} 2/3 & 1/3 \\ 1/3 & 2/3 \end{pmatrix}$$
$$M^2 = \begin{pmatrix} 5/9 & 4/9 \\ 4/9 & 5/9 \end{pmatrix}$$
$$M^3 = \begin{pmatrix} 14/27 & 13/27 \\ 13/27 & 14/27 \end{pmatrix}$$
$$M^4 = \begin{pmatrix} 41/81 & 40/81 \\ 40/81 & 41/81 \end{pmatrix},$$

and so on. What do we see? We notice that, as t increases, the entries of M^t get closer and closer to $1/2$. In fact, we can prove that, for this particular matrix M,

$$\lim_{t\to\infty} M^t = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}, \tag{1.25}$$

which means that, for any pair of states $x, y \in \{0,1\}$, $\lim_{t\to\infty} M^t(x,y) = 1/2$. Let's see what this means for p_t. Since X_t takes only two values, 0 and 1, it suffices to examine $p_t(0)$. Using Eqs. (1.15) and (1.25), we can write

$$\begin{aligned} \lim_{t\to\infty} p_t(0) &= \lim_{t\to\infty} \Big(p_0(0) M^t(0,0) + p_0(1) M^t(1,0) \Big) \\ &= p_0(0) \lim_{t\to\infty} M^t(0,0) + p_0(1) \lim_{t\to\infty} M^t(1,0) \\ &= \frac{1}{2} p_0(0) + \frac{1}{2} p_1(0) \\ &= \frac{1}{2}. \end{aligned} \tag{1.26}$$

Note that this holds regardless of the initial distribution p_0, so what Eq. (1.26) tells us is that, in the long run, both values of the state X_t of our two-state Markov chain will be equally likely. In a more formal mannner, let us denote by π the uniform distribution on $\{0,1\}$: $\pi = (1/2, 1/2)$. Then

$$\lim_{t\to\infty} p_t = \lim_{t\to\infty} p_0 M^t = \pi \tag{1.27}$$

for *any* initial distribution p_0. Moreover, it can be shown (and later on we will learn how to do that) that, in this particular example, the convergence in (1.27) happens exponentially quickly: there exists some positive constant $\theta < 1$, such that

$$|p_t(0) - \pi(0)| = \left| \pi_t(0) - \frac{1}{2} \right| \le \theta^t, \qquad \forall t \in \mathbb{Z}_+.$$

Now, let us note another curious fact: a simple calculation shows that $\pi M = \pi$. That is, if the initial state distribution π_0 is already uniform, it will remain uniform for all eternity. We say in this case that π is the *invariant* (or *equilibrium*) distribution of the Markov chain.

Now, any given Markov chain may have one equilibrium distribution, infinitely many equilibrium distributions, or no equilibrium distributions. We have already seen an example of the first possibility. To illustrate the second possibility, consider any finite-state Markov chain whose transition probability matrix is just the identity matrix, i.e.,

$$M(x,y) = \delta(x,y) \stackrel{\triangle}{=} \begin{cases} 1, & \text{if } x = y \\ 0, & \text{otherwise} \end{cases}.$$

In this case, for any distribution p on X we will have $pM = p$. As for an example of a chain with no equilibrium distributions, we revisit our good old symmetric random walk on the integers. Suppose that there exists an invariant distribution π. Let $\Pi(z)$ be its probability generating function. Then, from (**??**) we see that, for any z where $\Pi(z)$ exists and is finite, we must have

$$\Pi(z) = \left(\frac{1}{2}z^{-1} + \frac{1}{2}z\right)\Pi(z).$$

In particular, for any z such that $\Pi(z) \neq 0$ it must be the case that $\frac{1}{2}z^{-1} + \frac{1}{2}z = 1$. This equality, however, holds only when $z = 1$, so we arrive at a contradiction.

1.7 Markov chains in disguise: The PageRank algorithm

Markov chains provide a nice, tractable modeling framework for a wide variety of natural and engineered stochastic systems, and we will have ample opportunities to delve into more details later now. However, they pop up in seemingly nonstochastic contexts as well, and many popular algorithms in signal processing and machine learning turn out to be Markov chains in disguise. We close our first look at Markov chains with an example of such an algorithm: the PageRank algorithm, made (in)famous by Sergey Brin and Larry Page.[2]

Here is our problem: We have a huge collection of webpages which we would like to rank in order of importance. Suppose that the pages are indexed and numbered according to some arbitrary ordering, so we can identify the collection of all pages with a finite but huge set $\mathsf{X} = \{1, 2, \ldots, n\}$ (as of today, the number of indexed Web pages is around 5 billion[3], so n is on the order of billions). A *ranking* is an assignment of a nonnegative number $r(i)$ to each page $i \in \mathsf{X}$, where larger values correspond to higher ranks.

What properties should such a ranking have? The idea of Brin and Page was to ignore content and focus on the Web itself, i.e., on the linking relationships between pages. Intuitively, the more important a webpage is, the more links will lead to it from other pages (you can think of each link leading to a page as an endorsement of that page). On the other hand, only links from other important pages should count, and, if there are too many outgoing links from some page, we should not take these endorsements too seriously. So, let's see if we can make this a bit more formal. Given a pair of pages i, j, we will write $i \to j$ if there is a link from page i to page j. This way, we can form

[2] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: bringing order to the Web." Technical Report, Stanford InfoLab, 1998. Available online at http://ilpubs.stanford.edu:8090/422/.

[3] Source: http://www.worldwidewebsize.com/

the *Web graph*: treat each page as a vertex, and draw an arrow from page i to page j whenever $i \to j$. Now, for every i let us define

$$d(i) \triangleq \#(\text{pages } j \text{ such that } i \to j).$$

In other words, d_i is the number of (distinct) outgoing links from page i. Now, one way of formalizing our intuition of what a good ranking should be is as follows: for each i,

$$r(i) = \sum_{j: j \to i} \frac{r(j)}{d(j)}, \qquad i = 1, 2, \ldots, n. \tag{1.28}$$

First of all, why does this capture our intuitive requirements for a good ranking? Well, Eq. (1.28) says the following: if we want to determine the rank of a given page i, we look at all pages j that have links to i. Any such page j will have its own ranking, and if that ranking is high, then that should contribute to the ranking of page i. On the other hand, if page j is too generous with outgoing links (i.e., its $d(j)$ is high), we should dampen its influence on the overall rank of page i. Mathematically, (1.28) shows that, *if* such a ranking exists, then it must be a solution of a huge system of linear equations.

To write this system down in a more compact form, let us define the following $n \times n$ matrix L:

$$L(i,j) \triangleq \begin{cases} \frac{1}{d(i)}, & \text{if } i \to j \\ 0, & \text{otherwise} \end{cases}. \tag{1.29}$$

With this definition in place, we see that we can rewrite (1.28) as

$$r(i) = \sum_{j=1}^{n} r(j) L(j,i), \qquad i = 1, 2, \ldots, n$$

so, if we arrange the rankings $r(i)$ into a row vector $r = (r(1), r(2), \ldots, r(n))$, we see that the desired ranking r should be a solution of

$$r = rL. \tag{1.30}$$

The question is, does (1.30) have a nontrivial solution (i.e., $r(i) \neq 0$ for at least one i), and, if so, does it have a *nonnegative* solution, i.e., one where $r(i) \geq 0$ for all i?

Let us take a closer look at the matrix L. We can make the following immediate observation: if $d(i) > 0$, then $\sum_j L(i,j) = 1$; on the other hand, if $d(i) = 0$, then $L(i,j) = 0$ for all j. Such rows of all zeroes correspond to "black hole" pages without any outgoing links, and, apart from these rows, L looks like a matrix of one-step transition probabilities of a Markov chain with state space X. So, we can try the following fix: Any time you end up at one of those pages without any outgoing links, you just hit a magic "back button" and end up at some fixed page, say, page 1 (of course, we assume that $d(1) > 0$). This amounts to replacing each zero row of $L(i,j)$ with $(1, 0, 0, \ldots, 0)$. Formally, let a be a column vector in $\mathbb{R}^n$, where

$$a(i) = \begin{cases} 0, & \text{if } d(i) > 0 \\ 1, & \text{if } d(i) = 0 \end{cases}$$

and let $e_1 = (1, 0, \ldots, 0)$. So, we replace the original matrix L with another matrix S, given by

$$S = L + ae_1. \tag{1.31}$$

This has the desired effect of replacing each zero row with $(1, 0, \ldots, 0)$ (prove it!). Also, since n is huge, flipping a few zeros to ones will not perturb things too much. Now, S is a bona fide Markov matrix: all entries $S(i, j)$ are nonnegative, and all rows sum to one: for each i, $\sum_j S(i, j) = 1$. So, let us redefine our problem (1.30) as follows: we seek a nonnegative solution r to

$$r = rS. \tag{1.32}$$

Now, we are looking for a nontrivial nonnegative solution, and, because (1.32) is a linear system, there is no loss in assuming that the coordinates of r sum to one: $\sum_i r(i) = 1$. To see this, let r be a nontrivial solution of (1.32) with $r(i) \geq 0$ for all i, and replace each $r(i)$ with $r(i)/\sum_j r(j)$. Then the new vector r will also be a solution. Any such vector can be thought of as a *probability distribution* on $\mathsf{X} = \{1, \ldots, n\}$. Thus, we reformulate our problem as follows: the ranking we seek is an *equilibrium distribution* of a Markov chain with transition matrix S.

In fact, this Markov chain reformulation can be interpreted in terms of a so-called *random surfer model*. Imagine a Web surfer who starts from some randomly chosen page and then clicks on one of the links uniformly at random. Mathematically, let X_t denote the index of the page visited by the surfer at time t. This page contains $d(i)$ outgoing links, and the surfer chooses one of those links uniformly at random (each choice has probability $1/d(i)$) and clicks on it. The rows $(1, 0, \ldots, 0)$ correspond to those pages that did not have any outgoing links in the original Web graph, but now all have a back button to page 1. Thus,

$$\mathbf{P}[X_{t+1} = j | X_t = i] = \begin{cases} \frac{1}{d(i)}, & \text{if } i \to j \\ 0, & \text{otherwise} \end{cases},$$

which is exactly $S(i, j)$. It can be shown using the Law of Large Numbers that, under the random surfer model, the ranking $r(i)$ of the ith page is the limiting long-term frequency of visits to page i:

$$r(i) \approx \lim_{t \to \infty} \frac{\#\{X_s = i, 0 \leq s < t\}}{t}.$$

If such an equilibrium distribution exists, how do we find it? Solving the linear system (1.32) with billions of equstions is out of the question. But, we can cook up an alternative procedure, inspired by the apparent connection to Markov chains: Start with an arbitrary (and, most likely, arbitrarily bad) guess $r_0 = (r_0(1), \ldots, r_0(n))$, a probability distribution on X. Think of it as an initial ($t = 0$) state distribution for our Markov chain. Then $r_1 = r_0 S$ is the distribution at time 1, $r_2 = r_0 S^2$ is the distribution at time 2, $\ldots$, $r_t = r_0 S^t$ is the distribution at time t. Then we hope that, as t gets large, r_t will converge to the desired equilibrium distribution. However, this is not guaranteed to happen quickly, or at all. So, we implement one more tweak to help things along: we pick a small constant $\alpha \in (0, 1)$[4] and replace S with another matrix:

$$G \triangleq (1 - \alpha)S + \frac{\alpha}{n} E, \tag{1.33}$$

[4] Brin and Page suggested taking $\alpha = 0.15$.

where E is the $n \times n$ matrix of all ones. This is shorthand for

$$G(i,j) = (1-\alpha)S(i,j) + \frac{\alpha}{n}, \qquad i,j \in \{1,\dots,n\}.$$

This is, again, a Markov matrix (why?), and, as we will see later in the course, this little tweak guarantees that G has a unique invariant distribution, and that $r_0 G^t$ converges to this invariant distribution exponentially fast, for *any* initial guess r_0. The Markov chain with one-step transition matrix G describes the random surfer model with *teleportation*: At each time t, with probability α the surfer gets bored from following all the links and chooses a completely random new starting page. Again, because G is different from S, the resulting r that solves $r = rG$ will be different from the solution of (1.30), but the error will be small because n is so large. So, in very broad strokes, the PageRank algorithm looks like this:

pick an arbitrary initial distribution r_0
for $t = 1$ **to** N
  $r_t = r_{t-1} G$
end for
$r \leftarrow r_N$

In fact, one can formulate a wide variety of complex engineering problems in this way, by showing that the desired solution is a fixed point of some huge linear system. Taken far enough, this thinking can lead to some interesting speculations.