APM 541: Stochastic Modelling in Biology
Discrete-time Markov Chains

Jay Taylor
Fall 2013
Outline

1. Motivation
2. Markov Processes
3. Markov Chains: Basic Properties
4. Asymptotics: Class Structure
5. Asymptotics: Hitting Times and Probabilities
6. Asymptotics: Stationary Distributions
7. Time Reversal
8. Hidden Markov Models
Stochastic Processes

Definition

A stochastic process is simply a collection of random variables \( \{X_t : t \in T\} \) indexed by some set \( T \), where all of the variables are defined on the same underlying probability space. We say that the process is \( E \)-valued if all of the variables take values in the set \( E \). We say that the process is a discrete-time stochastic process if the index set \( T \) is a discrete subset of \( \mathbb{R} \), e.g., \( T \subset \mathbb{Z} \), or a continuous-time stochastic process if the index set is an interval \( T = (a, b) \subset \mathbb{R} \).

Interpretation: Usually, \( T \subset \mathbb{R} \) is a subset of the real line and we think of the \( X_t \) as specifying the state of some random system at time \( t \in T \). Thus the stochastic process as a whole describes the noisy or random evolution of the system over time.
To motivate our consideration of Markov processes, let us compare two models of density-dependent population regulation in a finite population of an asexual (or hermaphroditic) species. In both models, we will assume that reproduction is seasonal and we will let $X_t$ denote the number of individuals alive at the beginning of year $t$.

Model 1 will make the following assumptions:

- Individuals become reproductively mature in the year following their birth, i.e., an individual born in year $t$ will begin to reproduce in year $t + 1$.
- Conditional on the current population size $X_t$, the number of surviving offspring born to an individual alive at time $t$ is Poisson-distributed with mean $r \exp(-\gamma_b X_t)$. Furthermore, the numbers of offspring born to different individuals are independent of one another.
- Following reproduction, each individual dies, independently of the rest, with probability $1 - (1 - \mu) \exp(-\gamma_d X_t)$.
Assume that the individuals alive at the beginning of time $t$ are labeled $i = 1, \cdots, X_t$ and let $\eta_{i,t}$ denote the number of offspring born to the $i$'th individual in that season. Since $\eta_{1,t}, \cdots, \eta_{X_t,t}$ are independent Poisson-distributed random variables, it follows that if we condition on $X_t = n$, then the total number of offspring born at time $t$ is also Poisson-distributed, with mean $n\gamma b^n$:

$$
\eta_t \equiv \sum_{i=1}^n \eta_{i,t} \sim \text{Poisson}(n\gamma b^n).
$$

Furthermore, under this same condition, the number of surviving adults $S_t$ is binomially-distributed with parameters $n$ and $(1 - \mu)e^{-\gamma d^n}$:

$$
S_t \sim \text{Binomial}(n, (1 - \mu)e^{-\gamma d^n}).
$$

Combining these two observations, it follows that the number of individuals alive at the beginning of time $t + 1$ is

$$
X_{t+1} = S_t + \eta_t.
$$
This process can be simulated using the following procedure:

1. Ask the user to input values for the parameters $r, \mu, \gamma_b, \gamma_d$ as well as the initial number of individuals $X_0$ alive at time $t = 0$.

2. Given that $X_t = n$, generate independent random variables
   \[ \eta_t \sim \text{Poisson} \left( nre^{-\gamma_b n} \right) \] and \[ S_t \sim \text{Binomial} \left( n, (1 - \mu)e^{-\gamma_d n} \right). \]

3. Set $X_{t+1} = S_t + \eta_t$.

4. Increase $t$ to $t + 1$ and return to step 2.

Several features of this model contribute to the ease with which the process can be simulated. One is that we are able to directly sample the total number of individuals born in each season, rather than having to sample each of the individual offspring numbers separately. Equally importantly, to determine the number of individuals alive at time $t + 1$, we only need to know the number alive at time $t$. 
In the second model, we will assume that the density of the population at time $t$ affects not only the survival and reproduction of adults alive at that time, but also the development of any juveniles that are born in that year. Specifically, we will assume that density has a lasting negative effect on the future reproduction and survival of such juveniles, as described below:

1. Suppose that an individual is $\tau$ years old. Conditional on the population density $X_t$ in the present and on the population density $X_{t-\tau}$ in the year when that individual was born, the number of surviving offspring born to that individual at time $t$ is Poisson-distributed with mean

$$re^{-\gamma_b X_t} e^{-\lambda_b X_{t-\tau}}.$$

2. Following reproduction, the probability that this individual dies is

$$1 - \mu e^{-\gamma_d X_t} e^{-\lambda_d X_{t-\tau}}.$$

Apart from these modifications, we will assume that Model 2 coincides with Model 1.
As innocuous as these modifications may seem, they make the task of simulating Model 2 much more computationally demanding than Model 1. Specifically, we must deal with the following complications:

- Because reproduction and survival now depend on the density of the population in each individual’s natal year, we must keep track of the ages of each individual in the population so that we can determine when they were born. This also necessitates sampling Poisson and binomially-distributed random variables for every age class rather than for the population as a whole.

- More seriously, we must know the population density not only at the present $t$, but potentially at every time $0, 1, 2, \cdots t - 1$ in the past. In other words, the entire history of the population is needed to be able to propagate the process forward from time $t$ to $t + 1$.

The history-dependence exhibited by Model 2 is a major obstacle both to analytical and Monte Carlo investigations. Although this may sometimes be unavoidable, it is usually preferable to begin with a model that can be propagated forward using only the present state of the system.
Stochastic models that evolve in a manner that depends only on the current state are known as Markov processes. The formal definition is given below.

**Definition**

A stochastic process \( X = (X_n : n \geq 0) \) with values in a set \( E \) is said to be a **discrete time Markov process** if for every \( n \geq 0 \) and every set of values \( x_0, x_1, \cdots, x_n \in E \), we have

\[
P(X_{n+1} \in A | X_0 = x_0, X_1 = x_1, \cdots, X_n = x_n) = P(X_{n+1} \in A | X_n = x_n),
\]

whenever \( A \) is a subset of \( E \) such that \( \{X_{n+1} \in A\} \) is an event. In this case, the functions defined by

\[
p_n(x, A) = P(X_{n+1} \in A | X_n = x)
\]

are called the **one-step transition probabilities** of \( X \). If the functions \( p_n(x, A) \) do not depend on \( n \), i.e., if there is a function \( p \) such that

\[
p(x, A) = P(X_{n+1} \in A | X_n = x)
\]

for every \( n \geq 0 \), then we say that \( X \) is a **time-homogeneous Markov process** with **transition function** \( p \). Otherwise, \( X \) is said to be **time-inhomogeneous**.
Remarks:

- Markov processes are sometimes said to lack memory in the sense that if we know the current state of the process $X_t = x$, then the values assumed by the process at future times, say $X_{t+s} = y$, do not depend on how the process arrived at state $x$ at time $t$. In other words, the process is only aware of its current location and forgets how it arrived there.

- Equivalently, it can be shown that a Markov process has the following property, which is known as the Markov property. If we condition on the event $\{X_n = x\}$, then the variables $(X_{n+k} : k \geq 1)$ are independent of the variables $(X_{n-k} : k \geq 1)$, i.e., the future is conditionally independent of the past given the present.

- Although time-inhomogeneous processes have useful applications, here we will assume that all of our processes are time-homogeneous. Indeed, in some cases, we can convert a time-inhomogeneous process to one that is time-homogeneous by explicitly modeling the variables that are responsible for the changing transition probabilities.
Example: Discrete-time Random Walks

Let $\xi_0, \xi_1, \cdots$ be an i.i.d. sequence of $\mathbb{R}^d$-valued random variables with density $f$ and define the process $X = (X_n; n \geq 0)$ by setting $X_0 = 0$ and

$$X_{n+1} = X_n + \xi_n.$$ 

$X$ is said to be a $d$-dimensional discrete-time random walk and a simple calculation shows that $X$ is a time-homogeneous Markov process with transition function

$$
\mathbb{P}(X_{n+1} \in A|X_0 = x_0, \cdots, X_n = x_n) = \mathbb{P}(X_{n+1} \in A|X_n = x_n) \\
= \mathbb{P}(x_n + \xi_n \in A|X_n = x_n) \\
= \mathbb{P}(\xi_n \in A - x_n) \\
= \int_A f(y - x_n) \, dy.
$$

Interpretation: We think of $X_n$ as the location at time $n$ of an individual animal or particle that moves by taking random jumps at fixed time intervals.
Markov Chains

Markov processes that take values in a discrete state space $E$ are important enough to merit their own theory and terminology. In particular, it will be convenient for us to adopt the following two conventions:

1. **Without loss of generality, we will assume that the state space is either** $E = \{1, \cdots, m\}$ when $E$ is finite or $E = \{1, 2, \cdots\}$ when $E$ is countably infinite. This amounts to assigning integer labels to the original objects that were in the state space.

2. **Probability distributions on $E$ will be represented by row vectors, e.g., we will use the vector** $\nu = (\nu_1, \nu_2, \cdots)$ to represent the distribution that assigns probability $\nu_i$ to state $i \in E$. If $E$ is infinite, then we will regard $\nu$ as an infinite-dimensional vector.
With these conventions in force, we have the following definition.

**Definition**

A stochastic process \( X = (X_n; n \geq 0) \) with values in the countable set \( E = \{1, 2, \cdots \} \) is said to be a **time-homogeneous discrete-time Markov chain with initial distribution** \( \nu \) and **transition matrix** \( P = (p_{ij}) \) if

1. for every \( i \in E \), \( \mathbb{P}(X_0 = i) = \nu_i; \)
2. for every \( n \geq 0 \) and every set of values \( x_0, \cdots, x_{n+1} \in E \), we have

\[
\mathbb{P}(X_{n+1} = x_{n+1} | X_0 = x_0, X_1 = x_1, \cdots, X_n = x_n) = \mathbb{P}(X_{n+1} = x_{n+1} | X_n = x_n) = p_{x_n x_{n+1}}.
\]

**Caveat:** Some authors (in particular, demographers) define the transition matrix of a Markov chain to be the transpose of the matrix \( P \) that we have introduced above. In this case, \( p_{ji} \) is the probability that process transits from state \( i \) to state \( j \) in a single step.
Simulation of Discrete-time Markov Chains

In principle, a discrete-time Markov chain \( X = (X_n : n \geq 0) \) with initial distribution \( \nu \) and transition matrix \( P \) can be simulated using the following algorithm.

1. First, choose the initial state of the Markov chain \( X_0 = x_0 \) by sampling the value \( x_0 \) from the distribution \( \nu \) on \( E \).
2. If \( X_t = i \in E \), choose the next state of the Markov chain by sampling a value \( j \in E \) from the probability distribution \( (p_{i1}, p_{i2}, \cdots) \) on \( E \). Set \( X_{t+1} = j \).
3. Increase \( t \) to \( t + 1 \) and return to step 2.

In practice, the ease with which we can simulate a Markov chain depends on how difficult it is to generate random samples from the initial distribution \( \nu \) and from the transition probabilities specified by the rows of the transition matrix \( P \).
Example: Seasonal Fecundity in Birds

Etterson et al. (2009) propose a simple Markov chain model to describe the reproductive activities of a single female bird in a single breeding season. In these models, a female can occupy one of four states:

- state 1: she may be actively nesting
- state 2: she may have successfully fledged a brood
- state 3: she may have failed to fledge a brood
- state 4: she may have completed all nesting activities for the season.

Thus, the state space can be represented as $E = \{1, 2, 3, 4\}$, where each number corresponds to one of the four states described above, and the variable $X_n \in E$ describes the state of the female following the $n$'th change in state.
To complete the description of the Markov chain $X = (X_t : t \geq 0)$, we need to specify both the initial distribution $\nu = (1, 0, 0, 0)$ of the female’s state at the beginning of the season (i.e., she will initiate breeding) as well as the transition matrix $P = (p_{ij})$ determining how this state changes over time. Etterson and colleagues assume that the transition matrix can be written in the following form:

$$
P = \begin{pmatrix}
0 & s^a & 1 - s^a & 0 \\
1 - q_s & 0 & 0 & q_s \\
1 - q_f & 0 & 0 & q_f \\
0 & 0 & 0 & 1
\end{pmatrix} = \begin{pmatrix}
0 & 0.369 & 0.631 & 0 \\
0.33 & 0 & 0 & 0.67 \\
0.58 & 0 & 0 & 0.42 \\
0 & 0 & 0 & 1
\end{pmatrix}
$$

where $s$ is the daily nest survival probability, $a$ is the average time from first egg to fledging, $q_s$ is the probability that a female quits breeding following a successful breeding attempt, and $q_f$ is the probability that a female quits breeding following a failed breeding attempt. The numbers in the matrix on the right are estimates obtained from field studies of a population of Eastern Meadowlarks in Illinois.
The Eastern Meadowlark (*Sturnella magna*) is a member of the blackbird family (Icteridae).

Medium sized (length 7.5 - 10”), with a streaked brown back and yellow underparts with a black “V” on the breast.

Most common in grasslands, prairies and agricultural settings.

Eats mainly insects, as well as some seeds.
The matrices that occur as transition matrices of Markov chains are sufficiently important in their own right to warrant the following definition.

**Definition**

A matrix $P = (p_{ij})$, indexed by a countable set $E$, is said to be a **stochastic matrix** if all of the entries $p_{ij} \geq 0$ are non-negative and all of the row sums are equal to one, i.e.,

$$\sum_{j \in E} p_{ij} = 1$$

for every $i$.

Every transition matrix is a stochastic matrix, and it can be shown that every stochastic matrix is the transition matrix of some Markov chain. Furthermore, stochastic matrices satisfy a number of useful properties that are stated in the next lemma.
Lemma

1. **The product of any two stochastic matrices, \( P \) and \( Q \), of the same dimension is itself a stochastic matrix.**

2. **Every stochastic matrix \( P \) has \( \lambda = 1 \) as an eigenvalue, with corresponding right eigenvector \( v = (1, 1, \cdots) : Pv = v \).**

3. **Every eigenvalue \( \lambda \) of a stochastic matrix lies in the unit disk: \(|\lambda| \leq 1\).**

**Proof:** (1) Suppose that \( P = (p_{ij}) \) and \( Q = (q_{ij}) \) are stochastic matrices. Since all of the elements \( p_{ij} \) and \( q_{ij} \) are non-negative, it is evident that every element of \( PQ \),

\[
(PQ)_{ij} = \sum_{k \in E} p_{ik} q_{kj} \geq 0,
\]

is non-negative. Furthermore, every row sum of \( PQ \) is equal to 1 since

\[
\sum_{j \in E} (PQ)_{ij} = \sum_{j \in E} \sum_{k \in E} p_{ik} q_{kj} = \sum_{k \in E} p_{ik} \sum_{j \in E} q_{kj} = \sum_{k \in E} p_{ik} = 1.
\]
To establish (2), we simply need to show that $v = (1, 1, \cdots)$ is a right eigenvector for $P$ corresponding to eigenvalue 1. However, using the fact that each of the row sums of $P$ is equal to 1, we have

$$(Pv)_i = \sum_{k \in E} p_{ik} v_k = \sum_{k \in E} p_{ik} = 1,$$

which shows that $Pv = v$.

Lastly, to establish (3), we can either invoke Gershgorin’s disk theorem (look it up!) or observe that if $v = (v_1, v_2, \cdots)$ is any column vector with finite sup norm

$$||v|| \equiv \sup_{i \in E} |v_i| < \infty,$$

then

$$|(Pv)_i| = \left| \sum_{k \in E} p_{ik} v_k \right| \leq \sum_{k \in E} p_{ik} |v_k| \leq \sum_{k \in E} p_{ik} ||v|| = ||v||.$$

This shows that $||Pv|| \leq ||v||$ and so there can be no vector $v$ with $Pv = \lambda v$ when $|\lambda| > 1$. \qed
One consequence of the Markov property is that the joint distribution of the values assumed by a Markov chain from time 0 through time $n$ takes a particularly simple form.

**Theorem**

Let $X$ be a time-homogeneous discrete time Markov chain with transition matrix $P = (p_{ij})$ and initial distribution $\nu$ on $E$. Then

$$
\mathbb{P}(X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) = \nu(x_0) \prod_{i=0}^{n-1} p_{x_i, x_{i+1}}.
$$
Proof: By repeated conditioning and use of the Markov property, we have

\[
P(X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) =
\]

\[
= P(X_0 = x_0, \ldots, X_{n-1} = x_{n-1}) \cdot P(X_n = x_n | X_0 = x_0, \ldots, X_{n-1} = x_{n-1})
\]

\[
= P(X_0 = x_0, \ldots, X_{n-1} = x_{n-1}) \cdot P(X_n = x_n | X_{n-1} = x_{n-1})
\]

\[
= P(X_0 = x_0, \ldots, X_{n-1} = x_{n-1}) \cdot p_{x_{n-1}, x_n}
\]

\[
= \ldots
\]

\[
= P(X_0 = x_0, X_1 = x_1) \cdot p_{x_1, x_2} \cdots p_{x_{n-1}, x_n}
\]

\[
= P(X_0 = x_0) \cdot P(X_1 = x_1 | X_0 = x_0) \cdot p_{x_1, x_2} \cdots p_{x_{n-1}, x_n}
\]

\[
= \nu(x_0) \prod_{i=0}^{n-1} p_{x_i, x_{i+1}},
\]

where \( \nu(x_0) \) is the probability of \( x_0 \) under the initial distribution \( \nu \). \( \square \)
Example: Recall that Etterson et al. (2009) estimated the following transition matrix for the breeding cycle of Eastern Meadowlarks:

$$P = \begin{pmatrix}
0 & 0.369 & 0.631 & 0 \\
0.33 & 0 & 0 & 0.67 \\
0.58 & 0 & 0 & 0.42 \\
0 & 0 & 0 & 1
\end{pmatrix}.$$ 

Here, state 1 corresponds to active breeding, state 2 corresponds to fledging, state 3 corresponds to nesting failure, and state 4 corresponds to cessation of breeding activities in that season. If we assume that all females begin the season with an active breeding attempt ($X_0 = 1$), then the probability that a female successfully rears her first brood, fails to fledge her second brood and then successfully rears a third brood before stopping in the season is:

$$\mathbb{P}(1 \rightarrow 2 \rightarrow 1 \rightarrow 3 \rightarrow 1 \rightarrow 2 \rightarrow 4) = \nu_1 p_{12} p_{21} p_{13} p_{31} p_{12} p_{24} \approx 0.01102.$$
Multi-step Transition Probabilities

While the transition matrix specifies how a Markov chain will evolve from one time period to the next, for some applications we need to be able to calculate the transition probabilities over multiple time steps. As the name suggests, the \( n \)-step transition probabilities \( p_{ij}^{(n)} \) of a DTMC \( X \) are defined for any \( n \geq 1 \) by

\[
p_{ij}^{(n)} = \mathbb{P}(X_n = j | X_0 = i).
\]

In fact, our next theorem will show that the \( n \)-step transition probabilities are independent of time whenever \( X \) is time-homogeneous, i.e., for every \( t \geq 0 \),

\[
p_{ij}^{(n)} = \mathbb{P}(X_{t+n} = j | X_t = i),
\]

which means that \( p_{ij}^{(n)} \) is just the probability that the chain moves from state \( i \) to state \( j \) in \( n \) time steps. The next theorem expresses an important relationship that holds between the \( n \)-step transition probabilities of a DTMC \( X \) and its \( r \)- and \( n - r \)-step transition probabilities.
**Theorem**

*(Chapman-Kolmogorov Equations)* Assume that $X$ is a time-homogeneous discrete time Markov chain with $n$-step transition probabilities $p_{ij}^{(n)}$. Then, for any non-negative integer $r < n$, the identities

$$p_{ij}^{(n)} = \sum_{k \in E} p_{ik}^{(r)} p_{kj}^{(n-r)}$$

hold for all $i, j \in E$.

**Proof:** By using first the law of total probability and then the Markov property, we have

$$p_{ij}^{(n)} = \mathbb{P}(X_n = j | X_0 = i) = \sum_{k \in E} \mathbb{P}(X_n = j, X_r = k | X_0 = i)$$

$$= \sum_{k \in E} \mathbb{P}(X_n = j | X_r = k, X_0 = i) \cdot \mathbb{P}(X_r = k | X_0 = i)$$

$$= \sum_{k \in E} \mathbb{P}(X_n = j | X_r = k) \cdot \mathbb{P}(X_r = k | X_0 = i)$$

$$= \sum_{k \in E} p_{ik}^{(r)} p_{kj}^{(n-r)}.$$
Many computations involving Markov chains can be solved using techniques from linear algebra. To understand why this is so, let $P^{(n)} = (p_{ij}^{(n)})$ be the matrix containing the $n$-step transition probabilities and observe that the Chapman-Kolmogorov equations can be expressed as the following matrix equation:

$$P^{(n)} = P^{(r)} P^{(n-r)}.$$ 

In particular, if we take $n = 2$ and $r = 1$, then since $P^{(1)} = P$, we see that

$$P^{(2)} = PP = P^2.$$ 

This, in turn, implies that $P^{(3)} = PP^{(2)} = PP^2 = P^3$ and continuing in this fashion we find that

$$P^{(n)} = P^n.$$ 

In other words, the matrix of $n$-step transition probabilities is equal to the $n$’th power of the one-step transition matrix. This leads to the following formula for the marginal distribution of a Markov chain at time $n$. 

Theorem

Suppose that $X$ is a time-homogeneous discrete time Markov chain with transition matrix $P$ and initial distribution $\nu$. Then the distribution of $X_n$ is given by the column vector $\nu P^n$.

Proof: The result follows from the law of total probability:

$$
P(X_n = j) = \sum_{i \in E} P(X_n = j | X_0 = i) \cdot P(X_0 = i)
$$

$$
= \sum_{i \in E} \nu_i p_{ij}^{(n)}
$$

$$
= \sum_{i \in E} \nu_i \left( P^{(n)} \right)_{ij}
$$

$$
= (\nu P^n)_j.
$$
**Example:** To find the probability that a female meadowlark initiates at least 4 breeding attempts in a single season, we need to calculate the probability of the event \( X_6 = 1 \). Since the initial distribution is \( \nu = (1, 0, 0, 0) \), the distribution of \( X_6 \) is given by

\[
\nu P^6 = (1, 0, 0, 0) \begin{pmatrix} 0.116 & 0 & 0 & 0.884 \\ 0 & 0.029 & 0.0495 & 0.9215 \\ 0 & 0.0509 & 0.0871 & 0.862 \\ 0 & 0 & 0 & 1 \end{pmatrix}
\]

\[
= (0.116, 0, 0, 0.884).
\]

Thus, the probability that a female initiates breeding at least four times is equal to 0.116, while the probability that she initiates breeding three or fewer times is 0.884.
Long-term Behavior of Markov Chains

When studying stochastic models of biological processes, we are often interested in the long-term behavior of the model. Such considerations are important, for example, in the following questions:

- What is the probability that a new mutation will spread in a population?
- What is the probability that a tumor cell will escape immune surveillance and grow to form a tumor?
- What is the probability that an infectious disease will persist in a population?
- What is the steady-state distribution of transcription factors and other proteins within a single cell?

In the following sections, we will be concerned with two kinds of asymptotic behavior: absorption into a single state and long-term fluctuations in an equilibrium process.
We begin by introducing some useful terminology.

**Definition**

Let $X$ be a discrete time Markov chain on $E$ with transition matrix $P$.

1. We say that $i$ leads to $j$, written $i \rightarrow j$, if for some integer $n \geq 0$

$$p_{ij}^{(n)} = P_i (X_n = j) > 0.$$ 

In other words, $i \rightarrow j$ if the process $X$ beginning at $X_0 = i$ has some positive probability of eventually arriving at $j$.

2. We say that $i$ communicates with $j$, written $i \leftrightarrow j$, if $i$ leads to $j$ and $j$ leads to $i$. 
Notice that the relation $i \leftrightarrow j$ is an **equivalence relation** on $E$.

1. Each element communicates with itself: $i \leftrightarrow i$;
2. $i$ communicates with $j$ if and only if $j$ communicates with $i$;
3. If $i$ communicates with $j$ and $j$ communicates with $k$, then $i$ communicates with $k$.

In other words, a Markov chain with values in a state space $E$ induces a partition of $E$ into disjoint subsets $E_1, E_2, \ldots$,

$$E = E_1 \cup E_2 \cup \cdots$$

where all of the elements contained in the subset $E_i$ are communicating with each other and no two elements contained in different subsets, say $E_i$ and $E_j$, communicate with each other. This structure is formalized in the following definition.
Definition

Let $X$ be a DTMC on $E$ with transition matrix $P$.

1. A nonempty subset $C \subseteq E$ is said to be a communicating class if it is an equivalence class under the relation $i \leftrightarrow j$. In other words, each pair of elements in $C$ is communicating, and whenever $i \in C$ and $j \in E$ are communicating, $j \in C$.

2. A communicating class $C$ is said to be closed if whenever $i \in C$ and $i \rightarrow j$, we also have $j \in C$. If $C$ is a closed communicating class for a Markov chain $X$, then that means that once $X$ enters $C$, it never leaves $C$.

3. A state $i$ is said to be absorbing if $\{i\}$ is a closed class, i.e., once the process enters state $i$, it is trapped there forever.

4. A Markov chain is said to be irreducible if the entire state space $E$ is a communicating class.
Example: The transition matrix for the Markov chain studied by Etterson and colleagues is:

\[
P = \begin{pmatrix}
0 & sa & 1 - sa & 0 \\
1 - qs & 0 & 0 & qs \\
1 - qf & 0 & 0 & qf \\
0 & 0 & 0 & 1
\end{pmatrix}
\]

Provided that \( s \in (0, 1) \), \( a > 0 \), and \( qs, qf \in (0, 1) \), then the state space \( E = \{1, 2, 3, 4\} \) can be decomposed into two communicating classes, \( C_1 = \{1, 2, 3\} \) and \( C_2 = \{4\} \). This shows that the Markov chain is not irreducible. Furthermore, \( C_1 \) is not closed, since states 2 and 3 lead out of \( C_1 \) into \( C_2 \). On the other hand, \( C_2 \) is closed and 4 is an absorbing class.
Suppose that $X$ is a discrete time Markov chain and let $C \subset E$ be a closed communicating class for $X$. In this section we will consider the following two problems:

1. Given that $X_0 = i$, what is the probability that $X$ is eventually absorbed by $C$?
2. Given that $X_0 = i$, what is the average time to absorption in $C$?

Since we will be conditioning on the initial state of the Markov chain, it will be convenient to introduce the following notation. We will use $P_i$ to denote the conditional distribution of the chain given $X_0 = i$,

$$P_i(A) = P(A|X_0 = i),$$

where $A$ is any event defined in terms of the chain. Similarly, we will use $E_i$ to denote conditional expectations given $X_0 = i$,

$$E_i[Y] = E[Y|X_0 = i],$$

where $Y$ is any random variable defined in terms of the chain.
The main quantities of interest are defined below.

**Definition**

Let $X$ be a discrete time Markov chain on $E$ with transition matrix $P$ and let $C \subset E$ be a closed communicating class for $X$.

1. The **absorption time** of $C$ is the random variable $T^C \in \{0, 1, \cdots, \infty\}$ defined by

   $$T^C = \begin{cases} 
   \min \{n \geq 0 : X_n \in C\} & \text{if } X_n \in C \text{ for some } n \geq 0 \\
   \infty & \text{if } X_n \notin C \text{ for all } n.
   \end{cases}$$

2. The **absorption probability** of $C$ starting from $i$ is the probability

   $$h^C_i = \mathbb{P}_i \left( T^C < \infty \right).$$

3. The **mean absorption time** of $C$ starting from $i$ is the expectation

   $$\tau^C_i = \mathbb{E}_i \left[ T^C \right].$$
The following theorem asserts that the absorption probabilities $h_i^C$ can be calculated by solving a certain system of linear equations.

**Theorem**

The vector of absorption probabilities $h^C = (h_1^C, h_2^C, \cdots)$ is the unique minimal non-negative solution of the system of linear equations,

$$
\begin{cases}
    h_i^C = 1 & \text{if } i \in C \\
    h_i^C = \sum_{j \in E} p_{ij} h_j^C & \text{if } i \notin C
\end{cases}
$$

To say that $h^C$ is a minimal non-negative solution means that there is no other non-negative solution with at least one component smaller than that of $h^C$. In fact, it can be shown that the second set of equations is satisfied even when $i \in C$, which means that the vector $h^C$ is a right eigenvector for $P$ corresponding to the eigenvalue $\lambda = 1$:

$$h^C = Ph^C.$$
**Proof:** I will show that $h^C$ is a solution to this system of equations. For a proof of minimality, see the text *Markov Chains* by Norris (1997).

A strategy that is often successful when working with Markov chains is to condition on what the chain does in a single step. Here we will condition on the state of the chain at time $t = 1$. However, we first note that $h^C_i = 1$ if $i \in C$ since the chain will clearly absorb in $C$ if it starts there. On the other hand, if $i \notin C$, then the law of total probability and the Markov property imply that

\[
h^C_i \equiv \mathbb{P}_i (X_n \in C \text{ for some } n < \infty) = \sum_{j \in E} \mathbb{P}_i (X_1 = j \text{ and } X_n \in C \text{ for some } n < \infty) = \sum_{j \in E} \mathbb{P}_i (X_n \in C \text{ for some } n < \infty | X_1 = j) \cdot \mathbb{P}_i (X_1 = j) = \sum_{j \in E} \mathbb{P}_j (X_n \in C \text{ for some } n < \infty) \cdot p_{ij} = \sum_{j \in E} p_{ij} h^C_j.
\]
**Example:** Suppose that $X$ is a Markov chain on the state space $E = \{1, 2, 3, 4\}$ with the following transition matrix

$$
P = \begin{pmatrix}
1 & 0 & 0 & 0 \\
a & 0 & 1-a & 0 \\
0 & 1-b & 0 & b \\
0 & 0 & 0 & 1
\end{pmatrix},
$$

where $a, b \in (0, 1)$. Then both 1 and 4 are absorbing states and the chain can absorb in either one of these two states when it is started in state 2 or 3. Let $h_i$ be the probability that it eventually absorbs in state 4 when started in state $i$:

$$h_i = \mathbb{P}_i (X_n = 4 \text{ for some } n < \infty).$$

From the last theorem, we know that $h = (h_1, h_2, h_3, h_4)$ is the unique solution to the system of equations

$$h_i = \sum_{j \in E} p_{ij} h_j$$

with $h_4 = 1$. 

Written explicitly, this system of equations is:

\[
\begin{align*}
    h_1 &= h_1 \\
    h_2 &= ah_1 + (1-a)h_3 \\
    h_3 &= (1-b)h_2 + b,
\end{align*}
\]

which does not have a unique solution (since \( h_1 \) is unconstrained). However, if we insist on a minimal non-negative solution, then \( h_1 = 0 \), which in turn uniquely determines \( h_2 \) and \( h_3 \):

\[
\begin{align*}
    h_2 &= \frac{(1-a)b}{a + b - ab} \quad \text{and} \quad h_3 = \frac{b}{a + b - ab}.
\end{align*}
\]
The mean absorption times $\tau_i^C$ can also be calculated by solving a system of linear equations.

**Theorem**

Suppose that $C$ is a closed communicating class and that the absorption probability $h_i^C = 1$ for every $i \in E$, i.e., the chain is certain to absorb in $C$ no matter where started. Then, the vector of mean absorption times $\tau^C = (\tau_1^C, \tau_2^C, \cdots)$ is the minimal non-negative solution of the system of linear equations,

$$
\begin{align*}
\tau_i^C &= 0 & \text{if } i \in C \\
\tau_i^C &= 1 + \sum_{j \in E} p_{ij} \tau_j^C & \text{if } i \notin C
\end{align*}
$$

In this case, the system of equations is inhomogeneous, i.e., there are constant terms in the equations. Furthermore, the condition $h_i^C = 1$ is important, since otherwise some of the mean absorption times will be infinite (since we set $T^C = \infty$ if absorption never occurs).
Example: We can use this last result to calculate the mean number of breeding attempts by a female Eastern Meadowlark in the Etterson model. In this model, state 4 (cessation of breeding activities) is the unique absorbing state and the probability of absorption here is equal to 1. Using the transition matrix estimated by these authors, we obtain the following system of equations:

\[
\begin{align*}
\tau_1 &= 1 + 0.369 \tau_2 + 0.631 \tau_3 \\
\tau_2 &= 1 + 0.33 \tau_1 + 0.67 \tau_4 \\
\tau_3 &= 1 + 0.58 \tau_1 + 0.42 \tau_4 \\
\tau_4 &= 0.
\end{align*}
\]

This has a unique solution, which is \( \tau_1 \approx 3.904, \tau_2 \approx 2.288, \tau_3 \approx 3.264 \). These are the mean number of transitions that occur prior to absorption. To determine the mean number of breeding attempts, we take \( X_0 = 1 \) and observe that every breeding attempt passes through two events (active breeding followed by fledging or failure). Thus the mean number of breeding attempts is \( \tau_1/2 \approx 1.952 \).
Example: Birth-Death Processes

A discrete time Markov chain \( X = (X_n; \ n \geq 0) \) is said to be a birth-death process if \( X \) takes values in a set \( E = \{0, 1, \cdots, N\} \) with a tridiagonal transition matrix

\[
P = \begin{pmatrix}
1 - b_0 & b_0 & 0 & 0 & \cdots & 0 & 0 & 0 \\
0 & 1 - (b_1 + d_1) & b_1 & 0 & \cdots & 0 & 0 & 0 \\
0 & d_2 & 1 - (b_2 + d_2) & b_2 & \cdots & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & 0 & \cdots & d_{N-1} & 1 - (b_{N-1} + d_{N-1}) & b_{N-1} \\
0 & 0 & 0 & 0 & \cdots & 0 & d_N & 1 - d_N
\end{pmatrix},
\]

Such processes are often used to model the dynamics of populations in which individuals reproduce or die one at a time. However, a more general characterization of a birth-death process is that it is a Markov chain on a linearly ordered set in which transitions can only take place from a state to itself or to an immediately adjacent state. For example, birth-death processes can be used to model the movement of a motor protein such as dynein along a microtubule.
Suppose that a birth-death process $X$ is used to model a population of individuals that are reproducing and dying. In this case, $X_n$ will denote the number of individuals present in the population at time $n$. If $X_n = k$ with $1 \leq k \leq N - 1$, then there are three possibilities for $X_{n+1}$.

- With probability $b_k$, one of the $k$ individuals reproduces and the population size increases to $k + 1$. Here we are assuming that each reproduction results in the birth of exactly one offspring.
- With probability $d_k$, one of the $k$ individuals dies and the population size decreases to $k - 1$.
- With probability $1 - b_k - d_k$, no individual reproduces or dies during that time step and the population size remains at $k$. Alternatively, we could stipulate that this is the probability that exactly one individual gives birth and one individual dies, in which case there still is no change in the population size.
The boundary states 0 and $N$ need special consideration.

- If $b_0 = 0$, then 0 is an absorbing state corresponding to extinction of the population.
- If $b_0 > 0$, then 0 is no longer absorbing and we can think of this parameter as the probability (per unit time) that an extinct population is rescued by a migrant arriving from outside.
- If $N$ is finite, then $b_N = 0$ and we are tacitly assuming that density dependence is so strong that no individual reproduced when the population is this large. However, we can also take $N = \infty$, in which case the population may grow unboundedly.

Birth-death processes have been used extensively in the biological literature in part because they have many nice analytical properties that allow one to do explicit computations without having to resort to Monte Carlo simulations.
Suppose that $N$ is finite, that $b_0 = b_N = 0$, and that the remaining probabilities $b_1, \cdots, b_{N-1}$ and $d_1, \cdots, d_N$ are all positive. In this case 0 is the unique absorbing state for the Markov chain and it can be shown that the population is certain to go extinct at some finite time, which we will denote $T_{\text{extinct}}$. Our goal is to calculate the expected time to extinction as a function of the initial number of individuals in the population,

$$
\tau_k = \mathbb{E}_k \left[ T_{\text{extinct}} \right],
$$

where $k = 0, \cdots, N$.

By the preceding theorem, we know that the $\tau_k$’s satisfy the following system of equations:

$$
\begin{align*}
\tau_0 &= 0 \\
\tau_k &= 1 + d_k \tau_{k-1} + (1 - b_k - d_k)\tau_k + b_k \tau_{k+1}, \quad 1 \leq k \leq N - 1, \\
\tau_N &= 1 + d_N \tau_{N-1} + (1 - d_N)\tau_N.
\end{align*}
$$
By bringing all of the $\tau_k$’s over to the left-hand side, we can rewrite this system in the following more concise form:

\[
d_k(\tau_{k-1} - \tau_k) + b_k(\tau_{k+1} - \tau_k) = -1, \quad 1 \leq k \leq N - 1
\]
\[
d_N(\tau_{N-1} - \tau_N) = -1.
\]

Notice that this is a two-term recursion (i.e., $\tau_{k+2}$ depends only on $\tau_{k+1}$ and $\tau_k$), which is analogous two a second-order ODE and can be solved analogously, effectively by ‘integrating’ twice. For our first integration, we will define a new variable

\[
\Delta_k = \tau_{k+1} - \tau_k, \quad k = 0, \cdots, N - 1.
\]

Substituting this variable into the two equations above gives:

\[
b_k \Delta_k - d_k \Delta_{k-1} = -1, \quad 0 \leq k \leq N - 1
\]
\[
-d_N \Delta_{N-1} = -1.
\]
This latter system can be solved by starting at the upper boundary \((k = N - 1)\) and working down. Indeed, we have:

\[
\Delta_{N-1} = \frac{1}{d_N}
\]

\[
\Delta_{k-1} = \frac{b_k}{d_k} \Delta_k + \frac{1}{d_k}, \quad 0 \leq k \leq N - 1.
\]

This shows that

\[
\Delta_{N-2} = \frac{1}{d_{N-1}} + \frac{b_{N-1}}{d_{N-1} d_N}
\]

\[
\Delta_{N-3} = \frac{1}{d_{N-2}} + \frac{b_{N-2}}{d_{N-2} d_{N-1}} + \frac{b_{N-2} b_{N-1}}{d_{N-2} d_{N-1} d_N},
\]

with the general expression

\[
\Delta_k = \frac{1}{d_{k+1}} + \sum_{j=k+2}^{N} \frac{b_{k+1} \cdots b_{j-1}}{d_{k+1} \cdots d_j}
\]

for \(k = 0, \cdots, N - 1.\)
For our second ‘integration’, we will start at the lower boundary \( k = 0 \) and work up. Since \( \tau_0 = 0 \), we have

\[
\tau_1 = \tau_1 - \tau_0 = \Delta_0 = \frac{1}{d_1} + \sum_{j=2}^{N} \frac{b_1 \cdots b_{j-1}}{d_1 \cdots d_j}.
\]

Also,

\[
\begin{align*}
\tau_2 &= \tau_1 + \Delta_1 = \Delta_0 + \Delta_1 \\
\tau_3 &= \tau_2 + \Delta_2 = \Delta_0 + \Delta_1 + \Delta_2 \\
\cdots \\
\tau_k &= \sum_{i=0}^{k-1} \Delta_i
\end{align*}
\]

for \( k = 2, \cdots, N \). Using our result for \( \Delta_j \), it follows that the mean time to extinction when there are \( k \) individuals is

\[
\tau_k = \sum_{i=0}^{k-1} \left[ \frac{1}{d_{i+1}} + \sum_{j=i+2}^{N} \frac{b_{i+1} \cdots b_{j-1}}{d_{i+1} \cdots d_j} \right].
\]
Remarks:

- The results described in the previous two theorems are most useful when the state space is small or the transition matrix is sparse, i.e., most of the transition probabilities $p_{ij}$ are equal to 0. Indeed, sparseness is part of the reason that birth-death processes are so amenable to analytical calculations.

- If these conditions are not satisfied, then it may impractical to solve these systems of linear equations. In such cases, absorption probabilities and mean absorption times can be estimated by conducting stochastic simulations of the model and observing the proportion of simulations that end in absorption in a closed class $C$, as well as the times required to arrive in $C$. 
Stationary Distributions

Recall that we say that a Markov chain is **irreducible** if the entire state space is a communicating class, i.e., if every pair of states is communicating. In particular, because irreducible chains lack absorbing states, we usually cannot predict the states that will be occupied by the chain at distant times in the future, especially if the initial distribution is unknown. On the other hand, sometimes we can determine the distribution of the chain at such future times even when we don’t know the initial distribution. The following definition is key.

**Definition**

Let $X$ be a discrete time Markov chain with values in $E$ and transition matrix $P$. A distribution $\pi$ on $E$ is said to be a **stationary distribution** for $X$ if

$$\pi P = \pi.$$  

In other words, $\pi$ is a left eigenvector for $P$ with corresponding eigenvalue 1.
The probabilistic meaning of stationary distributions is revealed by the following theorem.

**Theorem**

Suppose that $\pi$ is a stationary distribution for a Markov chain $X$ with transition matrix $P$. If $\pi$ is the distribution of $X_0$, then $\pi$ is also the distribution of $X_n$ for every $n \geq 0$.

**Proof:** Since $X$ has initial distribution $\pi$ and transition matrix $P$, we know that the distribution of $X_n$ is

$$\pi P^n = (\pi P) P^{n-1} = \pi P^{n-1} = \cdots = \pi.$$ 

In other words, any stationary distribution of a Markov chain is time invariant: if ever the process has $\pi$ as its distribution, then it will retain this distribution at all future times.
Example: Suppose that $X$ is a Markov chain on the state space $E = \{1, 2\}$ with the following transition matrix

$$
P = \begin{pmatrix}
a & 1-a \\
1-b & b
\end{pmatrix},
$$

where $a, b \in (0, 1)$. Since $\pi = (p, 1-p)$ is a stationary distribution for $X$ if and only if $p$ satisfies the following system of equations,

\[
\begin{align*}
ap + (1-b)(1-p) &= p \\
(1-a)p + b(1-p) &= 1-p,
\end{align*}
\]

it follows that

$$
\pi = \left( \frac{1-b}{2-a-b}, \frac{1-a}{2-a-b} \right)
$$

is the unique stationary distribution for $X$. 
As the following examples illustrate, neither existence nor uniqueness of stationary distributions is assured.

**Example:** Let $Z_1, Z_2, \cdots$ be a sequence of i.i.d. Bernoulli random variables with success probability $p > 0$ and let $X = (X_n : n \geq 0)$ be the random walk defined by setting $X_0 = 0$ and

$$X_n = Z_1 + \cdots + Z_n, \quad n \geq 1.$$ 

By the strong law of large numbers, we know that

$$\mathbb{P}\left( \lim_{n \to \infty} \frac{1}{n} X_n = p \right) = 1.$$ 

However, this implies that $X_n$ tends to infinity almost surely and so $X$ has no stationary distribution on the integers.
Example: Suppose that \( X \) is a Markov chain on the state space \( E = \{1, 2, 3, 4\} \) with the following transition matrix

\[
P = \begin{pmatrix}
1 & 0 & 0 & 0 \\
0 & a & 0 & 1 - a \\
0 & 1 - b & 0 & b \\
0 & 0 & 0 & 1
\end{pmatrix},
\]

where \( a, b \in (0, 1) \). As we saw earlier, 1 and 4 are both absorbing states. Furthermore, any distribution of the form \((p, 0, 0, 1 - p)\) with \( p \in [0, 1] \) is a stationary distribution of the chain and so there are uncountably infinitely many stationary distributions. Indeed, any chain with more than one absorbing distribution will have infinitely many stationary distributions.
Under certain conditions, we can guarantee that not only will a Markov chain have a unique stationary distribution $\pi$, but also that the distribution of the chain at time $n$ will tend to $\pi$ as $n \to \infty$ no matter what the initial distribution may be. For this to be true, we need to exclude certain types of oscillatory behavior, hence the following definition.

**Definition**

A discrete time Markov chain with values in $E$ and transition matrix $P$ is said to be aperiodic if for every state $i \in E$ we have $p_{ii}^{(n)} > 0$ for all sufficiently large $n$.

The need for aperiodicity is illustrated by the following example. If $X$ is the Markov chain with transition matrix

$$P = \begin{pmatrix}
0 & 1 \\
1 & 0
\end{pmatrix},$$

then $\pi = (1/2, 1/2)$ is the unique stationary distribution. However, the distribution of the chain will be equal to the initial distribution at all future times $n$ with $n$ even. Indeed, this chain is periodic with period 2.
The following theorem provides sufficient conditions for a chain to converge to a unique stationary distribution from any initial distribution.

**Theorem**

Suppose that $X$ is an irreducible, aperiodic Markov chain with values in a finite set $E$ and transition matrix $P$. Then $X$ has a unique stationary distribution $\pi$ and for any initial distribution $\mu$, we have

$$\lim_{n \to \infty} P(X_n = j) = \pi_j$$

for every $j \in E$. In particular,

$$\lim_{n \to \infty} p_{ij}^{(n)} = \pi_j$$

for all $i,j \in E$.

**Remark:** Although existence of a stationary distribution is not guaranteed if we allow $E$ to be infinite, the uniqueness and convergence parts of this theorem remain true for arbitrary state spaces whenever a stationary distribution can be shown to exist.
Example: If $X$ is a Markov chain on the state space $E = \{1, 2\}$ with transition matrix

$$P = \begin{pmatrix} a & 1 - a \\ 1 - b & b \end{pmatrix},$$

where $a, b \in (0, 1)$, then $X$ is aperiodic and irreducible and

$$P^n = \begin{pmatrix} \frac{1-b}{2-a-b} & \frac{1-a}{2-a-b} \\ \frac{1-b}{2-a-b} & \frac{1-a}{2-a-b} \end{pmatrix} + (a + b - 1)^n \begin{pmatrix} \frac{1-a}{2-a-b} & -\frac{1-a}{2-a-b} \\ -\frac{1-b}{2-a-b} & \frac{1-b}{2-a-b} \end{pmatrix}$$

for every $n \geq 0$. Furthermore, since $|a + b - 1| < 1$, it is clear that

$$\lim_{n \to \infty} P^n = \begin{pmatrix} \frac{1-b}{2-a-b} & \frac{1-a}{2-a-b} \\ \frac{1-b}{2-a-b} & \frac{1-a}{2-a-b} \end{pmatrix},$$

where each row of the limiting matrix is the stationary distribution $\pi$ that we found by direct calculation. In particular, we see that the rate of convergence to the stationary distribution is geometric and depends on the magnitude of $a + b - 1$. 

Jay Taylor (ASU)  
APM 541  
Fall 2013  57 / 92
Convergence to Stationarity

It is often (although not always) the case that convergence to the stationary distribution occurs at a geometric rate. For example, let $P$ be the transition matrix of a Markov chain on a finite space $E = \{1, \cdots, n\}$ with stationary distribution $\pi$ and suppose that $P$ has $n$ distinct eigenvalues $\lambda_1, \cdots, \lambda_n$ with

$$1 = \lambda_1 \geq |\lambda_2| \geq \cdots \geq |\lambda_n|.$$  

In this case, $P$ is diagonalizable and there is a basis of left eigenvectors, say $\pi^{(1)}, \cdots, \pi^{(n)}$, with $\pi^{(1)} = \pi$, such that

$$\pi^{(i)} P = \lambda_i \pi^{(i)}.$$  

In particular, any distribution $\nu$ on $E$ can be expanded as a linear combination of left eigenvectors, say

$$\nu = c_1 \pi^{(1)} + \cdots + c_n \pi^{(n)}.$$
Suppose that $\nu$ is the initial distribution of the chain and that $|\lambda_2| < 1$. Then the distribution of $X_t$ is given by

$$
\nu P^t = (c_1 \pi^{(1)} + \cdots + c_n \pi^{(n)}) P^t \\
= c_1 \lambda_1^t \pi^{(1)} + c_2 \lambda_2^t \pi^{(2)} + \cdots + c_n \lambda_n^t \pi^{(n)} \\
= c_1 \pi + O(|\lambda_2|^t) \\
\to \pi
$$

as $t \to \infty$. This shows that $c_1 = 1$ and that $\nu P^t$ converges to $\pi$ geometrically at a rate which depends on the leading non-unit eigenvalue, $\lambda_2$. 
Example: Suppose that $X$ is a birth-death process with $N < \infty$ and that the probabilities $b_0, \cdots, b_{N-1}$ and $d_1, \cdots, d_N$ lie in the interval $(0, 1)$, i.e., $X$ is a birth-death process with immigration into extinct populations. In this case, $X$ is also an irreducible, aperiodic Markov chain on a finite state space and so $X$ has a unique stationary distribution $\pi$ which satisfies the following system of equations:

$$
(1 - b_0) \pi_0 + d_1 \pi_1 = \pi_0 \\
b_{k-1} \pi_{k-1} + (1 - b_k - d_k) \pi_k + d_{k+1} \pi_{k+1} = \pi_k \quad k = 1, \cdots, N - 1 \\
b_{N-1} \pi_{N-1} + (1 - d_N) \pi_N = \pi_N.
$$

By subtracting the probabilities on the right-hand side from both sides, this system can be rewritten in the following form:

$$
-b_0 \pi_0 + d_1 \pi_1 = 0 \\
b_{k-1} \pi_{k-1} - (b_k + d_k) \pi_k + d_{k+1} \pi_{k+1} = 0 \quad k = 1, \cdots, N - 1 \\
b_{N-1} \pi_{N-1} - d_N \pi_N = 0.
$$
Because each equation depends only on two or three successive probabilities, it is possible to solve this system recursively. The first equation can be rewritten as

\[ d_1 \pi_1 = b_0 \pi_0, \]

which implies that

\[ \pi_1 = \frac{b_0}{d_1} \pi_0. \]

Taking \( k = 1 \), we have

\[ b_0 \pi_0 - d_1 \pi_1 - b_1 \pi_1 + d_2 \pi_2 = 0. \]

However, since the first two terms cancel, this reduces to

\[ -b_1 \pi_1 + d_2 \pi_2 = 0. \]

which shows that

\[ \pi_2 = \frac{b_1}{d_2} \pi_1 = \frac{b_1 b_0}{d_2 d_1} \pi_0. \]
Continuing in this way, we find that

\[ \pi_k = \frac{b_{k-1}}{d_k} \pi_{k-1} = \left( \frac{b_{k-1} \cdots b_0}{d_k \cdots d_1} \right) \pi_0 \]

for \( k = 1, \cdots, N \). All that remains to be determined is \( \pi_0 \). However, since \( \pi \) is a probability distribution on \( E \), the probabilities must sum to one. Consequently,

\[ 1 = \sum_{k=0}^{N} \pi_k = \pi_0 \left( 1 + \sum_{k=1}^{N} \frac{b_{k-1} \cdots b_0}{d_k \cdots d_1} \right). \]

which shows that

\[ \pi_0 = \left( 1 + \sum_{k=1}^{N} \frac{b_{k-1} \cdots b_0}{d_k \cdots d_1} \right)^{-1}. \]
Combining these results, we see that the stationary distribution for this birth-death chain is

$$
\pi_k = \left( \frac{b_{k-1} \cdots b_0}{d_k \cdots d_1} \right) \left( 1 + \sum_{j=1}^{N} \frac{b_{j-1} \cdots b_0}{d_j \cdots d_1} \right)^{-1}
$$

for $k = 1, \cdots, N$, with $\pi_0$ as above. For example, if $b_0 = m$, $b_1 = \cdots = b_{N-1} = b$ and $d_1 = \cdots = d_N = d$, then

$$
\pi_k = \begin{cases} 
\left( 1 + \frac{m}{d} \frac{1-\gamma^N}{1-\gamma} \right)^{-1} & k = 0 \\
\frac{m}{d} \gamma^{k-1} \left( 1 + \frac{m}{d} \frac{1-\gamma^N}{1-\gamma} \right)^{-1} & k = 1, \cdots, N,
\end{cases}
$$

provided $\gamma \equiv b/d \neq 1$. 
Time Reversal of Markov Chains

Suppose that $X = (X_t : t \in \mathbb{N})$ is a discrete time Markov chain and for each $t > 0$, let 

$$\tilde{X}_t = X_{T-t}.$$ 

Then $\tilde{X} = (\tilde{X}_t : t \geq 0)$ is itself a Markov process on $E$, called the \textbf{time reversal of $X$}, and the transition probabilities of this Markov chain are given by

$$\tilde{p}_{xy}(t) \equiv \mathbb{P}(\tilde{X}_{t+1} = y | \tilde{X}_t = x) = \mathbb{P}(X_{T-(t+1)} = y | X_{T-t} = x)$$

$$= \mathbb{P}(X_{T-t} = x | X_{T-t-1} = y) \frac{\mathbb{P}(X_{T-t-1} = y)}{\mathbb{P}(X_{T-t} = x)}$$

$$= p_{yx} \frac{\mathbb{P}(X_{T-t-1} = y)}{\mathbb{P}(X_{T-t} = x)}.$$

As is clear from the last line, the transition probabilities of the time reversal of a Markov chain are usually time inhomogeneous even if the original chain is time homogeneous.
One setting in which the time reversal of a Markov chain is certain to be time homogeneous is when the original chain is stationary, say with distribution $\pi$. In this case, the distribution of $X_t$ is $\pi$ for all $t$ and so the transition probabilities of the time reversed process are

$$\tilde{p}_{xy} \equiv p_{yx} \frac{\mathbb{P}(X_{T-t-1} = y)}{\mathbb{P}(X_{T-t} = x)} = p_{yx} \frac{\pi_y}{\pi_x}.$$ 

Furthermore, if we let $\Pi$ be the diagonal matrix with entries $\pi$ along the diagonal (and zeros everywhere else), then the transition matrix $\tilde{P}$ of the time-reversed chain is

$$\tilde{P} = \Pi^{-1} P^T \Pi,$$

where $P^T$ denotes the transpose of $P$. 
It is usually the case that a stationary Markov chain and its time reversal have different transition matrices and therefore are fundamentally different processes. The one exception to this case occurs when the condition described in the following definition is satisfied.

**Definition**

Suppose that $X$ is a discrete time Markov chain on $E$ with transition matrix $P$ and let $\pi$ be a stationary distribution for $X$. The chain is said to satisfy the detailed balance conditions with respect to $\pi$ if the following identities are satisfied:

$$\pi_x p_{xy} = \pi_y p_{yx}$$

for every pair of states $x, y \in E$.

If the detailed balance conditions are satisfied, then

$$\tilde{p}_{xy} = \frac{p_{yx} \pi_y}{\pi_x} = \frac{p_{xy} \pi_x}{\pi_x} = p_{xy},$$

which shows that $\tilde{P} = P$, i.e., the chain and its time reversal have the same transition matrix and so the process ‘looks’ the same run forwards or backwards in time.
**Example:** Let $X$ be the birth-death process on $E = \{0, 1, \cdots, N\}$ with stationary distribution

$$
\pi_k = \begin{cases} 
(1 + \sum_{k=1}^{N} \frac{b_k \cdots b_0}{d_k \cdots d_1})^{-1} & k = 0 \\
\left(\frac{b_k \cdots b_0}{d_k \cdots d_1}\right) \left(1 + \sum_{k=1}^{N} \frac{b_k \cdots b_0}{d_k \cdots d_1}\right)^{-1} & k = 1, \cdots, N.
\end{cases}
$$

Then $X$ satisfies the detailed balance conditions with respect to $\pi$. Indeed, since most of the transition probabilities are zero, we need only check the following cases:

$$
\pi_0 p_{01} = \pi_0 b_0 \quad \pi_1 d_1 = \pi_1 p_{10} \\
\pi_k p_{k,k+1} = \pi_k b_k \quad \pi_{k+1} d_{k+1} = \pi_{k+1} p_{k+1,k}
$$

for $k = 1, \cdots, N$. This shows that the time reversal of a stationary birth-death process is also a birth-death process with the same birth and death probabilities.
Hidden Markov Models

Models and Observations

Suppose that we are interested in the spread of a communicable infectious disease through a population and that we decide to investigate this using the stochastic SIR model introduced previously:

\[
\begin{align*}
S_{t+1} &= S_t - w_t \\
I_{t+1} &= I_t + w_t - r_t \\
R_{t+1} &= R_t + r_t \\
w_t &\sim \text{Binomial} \left( S_t, 1 - (1 - \gamma + \gamma(1 - \beta))I_t \right) \\
r_t &\sim \text{Binomial}(I_t, \rho).
\end{align*}
\]

Here, \( w_t \) and \( r_t \) are the numbers of new infections and new recoveries between day \( t \) and \( t + 1 \), respectively, while \( \gamma \) is the probability (per day) of a contact between two individuals, \( \beta \) is the probability of disease transmission given that there is a contact between a susceptible and an infected individual, and \( \rho \) is the probability (per day) of recovery.
Since $S_t + I_t + R_t = N$ is constant for all $t \geq 0$, the variables $\{(S_t, I_t) : t \geq 0\}$ form a discrete-time Markov chain with values in the state space

$$E = \{(s, i) : i, s \geq 0, i + s \leq N\}$$

and transition probabilities

$$p_{(s,i),(s',i')} = \mathbb{P}\{(S_{t+1}, I_{t+1}) = (s', i')|(S_t, I_t) = (s, i)\}$$

$$= \binom{s}{s'} \pi_i^{s-s'} (1 - \pi_i)^{s'} \binom{i}{i' + s - i' - s'} \rho^{i'+s'-s} (1 - \rho)^{i' + s' - s}$$

where $\pi_i = 1 - (1 - \gamma + \gamma(1 - \beta))^i$ and we require that $0 \leq i - i' + s - s' \leq i$. While the transition matrix is complicated, at the very least we can conduct Monte Carlo simulations of this process to investigate quantities such as the duration of the epidemic and the total number of infections.
However, what if the values of the parameters $\gamma$, $\beta$ and $\rho$ are not known in advance, but instead must be estimated from data? There are two scenarios.

- **Scenario I:** Suppose that we have access to time series data consisting of the numbers of susceptible and infected individuals at each time in some period $[0, T]$, i.e., we have data $D = \{(s_t, i_t) : 0 \leq t \leq T\}$, and we believe this data to be accurate.

In this case, we can use the Markov property to calculate the probability of the data for each possible set of parameter values and we will call the resulting quantity the **likelihood** of the parameter values $\Theta = (\gamma, \beta, \rho)$

$$L(\theta|D) \equiv \mathbb{P}((S_t, l_t) = (s_t, i_t), 0 \leq t \leq T|\Theta)$$

$$= \prod_{t=0}^{T-1} p(s_t, i_t, (s_{t+1}, i_{t+1})(\Theta)).$$
The true values of the parameters can then be estimated either by finding the value \( \hat{\Theta}_{ML} \) that maximizes the likelihood function

\[
\hat{\Theta}_{ML} \equiv \arg \max_{\Theta} L(\Theta|D),
\]

or by choosing a prior distribution for the parameters, say \( p(\Theta) \), and then using Bayes’ formula in conjunction with the likelihood function to calculate the posterior distribution of the parameters given the data \( D \):

\[
p(\Theta|D) = p(\Theta) \frac{L(\Theta|D)}{P(D)},
\]

where

\[
P(D) = \int_{\Theta} L(\Theta|D) d\Theta = \int_{\Theta} P(D|\Theta) d\Theta
\]

is the so-called prior predictive probability of the data. The first approach is called maximum likelihood estimation, while the second approach is Bayesian inference, which we will discuss in some depth later in the semester.
**Scenario II:** Now suppose that many disease cases go undetected or unreported, but that we have access to time series data consisting of case reports at each time in the period \([0, T]\), i.e., our data consists of \(D = \{c_t : 0 \leq t \leq T\}\), where \(c_t\) is the number of reported infections that were contracted at time \(t\). (For convenience, we will assume that physicians are able to determine when an infected individual contracted the disease.)

To be able to use this data for parameter estimation, we need to extend the model so that it describes the joint probability distribution of both the state variables \(S_t\) and \(I_t\) and the observables \(C_t\). Here we will simply assume that each new infection has a fixed probability, say \(\delta\), of being detected and reported, in which case

\[ C_t \sim \text{Binomial}(I_t, \delta). \]

With this assumption, the extended process \(\{(S_t, I_t, C_t) : t \geq 0\}\) becomes a discrete-time Markov chain and we can write down the transition matrix and conduct Monte Carlo simulations. However, estimation of the model parameters from the data available to us remains challenging because the observables \(\{C_t : t \geq 0\}\) do not form a Markov process on their own.
To illustrate the difficulties posed by this data, suppose that we want to calculate the likelihood function using only the observed data:

\[
L(\Theta|D) = \mathbb{P} \left( \left( c_t \right)_{0:T}; \Theta \right) \\
= \sum_{\mathcal{H}(0:T)} \mathbb{P} \left( \mathcal{H}(0:T), \left( c_t \right)_{0:T}; \Theta \right),
\]

where \( \mathcal{H}(0:T) \) denotes a possible history for the state variables \( (S_t, I_t) \) between times 0 and \( T \). Although the latter expression is exact and each of the probabilities \( \mathbb{P} \left( \mathcal{H}(0:T), \left( c_t \right)_{0:T}; \Theta \right) \) is straightforward to compute using the fact that the extended process \( (S_t, I_t, C_t) \) is a Markov chain with a known transition matrix, we probably won’t be able to sum over all of the possible histories except when either \( N \) or \( T \) or both are quite small.

Fortunately, it is possible to carry out inference in cases like that described by Scenario II, but to do so we need to introduce a new class of stochastic processes called Hidden Markov models, which we now do.
Hidden Markov Models

Definition

A process \( \{(X_t, Y_t) : t \geq 0\} \) is said to be a Hidden Markov Model (HMM) if the following conditions are satisfied.

- The process \( X = (X_t : t \geq 0) \) is a discrete-time Markov chain with values in a state space \( S = \{s_1, s_2, \cdots\} \), initial distribution \( \mu \) and transition matrix \( P \).
- \( V = \{o_1, o_2, \cdots\} \) is a countable set and for each state \( s \in S \), \( e(\cdot|s) \) is a probability distribution on \( V \). \( e(o|s) \) is said to be the emission probability of observation \( o \) in state \( s \).
- Conditional on \( X_t = s \), the variable \( Y_t \) is independent of all of the remaining variables and
\[
P(Y_t = o | X_t = s, (X_u, Y_u)_{u \neq t}) = e(o|s).
\]

Informally, we should think of a hidden Markov model as consisting of an unobserved (hence hidden) Markov chain \( X \), as well as a sequence of observations \( Y_0, Y_1, Y_2, \cdots \) such that \( Y_t \) depends directly only on \( X_t \).
The statistical relationships between the variables in a hidden Markov model can be described with the help of a simple directed graph:

\[
\begin{align*}
Y_0 & \quad Y_1 & \quad Y_2 & \quad Y_3 & \quad Y_4 & \quad Y_5 \\
\uparrow e & \quad \uparrow e & \quad \uparrow e & \quad \uparrow e & \quad \uparrow e & \quad \uparrow e \\
X_0 & \quad p \rightarrow X_1 & \quad p \rightarrow X_2 & \quad p \rightarrow X_3 & \quad p \rightarrow X_4 & \quad p \rightarrow X_5
\end{align*}
\]

This diagram can be interpreted in the following way.

- Each vertex \( v \) corresponds to a particular variable \( Z_v \).
- \( Z_v \) is said to be a parent of a variable \( Z_{v'} \) if there is a directed edge from \( v \) to \( v' \).
- Any set of variables \( Z_{v_1}, \ldots, Z_{v_k} \) is conditionally independent given their parental variables \( Z_{pa(v_1)}, \ldots, Z_{pa(v_k)} \).

Because this last property holds, HMM’s belong to an even larger class of stochastic graphical models called **Bayesian networks**.
There are three basic problems that must be addressed when HMM’s are to be used:

- **Evaluation**: Given a sequence of observations, \( y_{0:t} \), how do we calculate the probability of that sequence?
- **Decoding**: Given a sequence of observations, \( y_{0:t} \), how can we make inferences about the unseen sequence of states \( x_{0:t} \) that generated these observations?
- **Estimation**: Given one or more sequences of observations, how can we estimate the parameters of the HMM?

**Notation**: I will often use the symbol \( x_{0:t} \) as short hand for the sequence \( (x_0, \ldots, x_t) \). Also, unless ambiguous, I will usually suppress the variables \( X_t, Y_t \) when writing probabilities of events involving the HMM, e.g.,

\[
P(x_{0:t}) \equiv P(X_0 = x_0, \ldots, X_t = x_t).
\]
Example: Flipping Biased Coins

Suppose that we play a game in which we are given two coins and at each stage we must toss one of them. Assume that the following are true:

- One coin is unbiased, but the other coin lands on heads with probability 0.7.
- We have no direct way of distinguishing between the two coins apart from observing the outcomes of the tosses.
- In each step, with probability 0.8, we toss the coin that we are currently holding. Otherwise, we exchange this coin for the other one and toss it.

This leads to a very simple HMM, in which the hidden Markov chain takes values in the set \{C_1, C_2\} and has transition matrix

\[
P = \begin{pmatrix}
0.8 & 0.2 \\
0.2 & 0.8
\end{pmatrix},
\]

while the emission probabilities are \(e(H|C_1) = e(T|C_1) = 0.5\) and \(e(H|C_2) = 0.7\) and \(e(T|C_2) = 0.3\).
Example: CpG Islands

- CpG dinucleotides are underrepresented in most parts of the human and other vertebrate genomes because of the fortuitous interaction of two processes.

- On the one hand, most CpG dinucleotides are epigenetically modified by the addition of a methyl (−CH$_3$) group to the cytosine to form 5-methylcytosine. These marks are used to distinguish parental and daughter DNA strands during replication and play a role in proofreading.

- On the other hand, the base 5-methylcytosine undergoes a relatively high rate of spontaneous deamination to the base thymine. This leads to a deficiency of CpG dinucleotides and an excess of TpG and CpA dinucleotides.

- CpG islands are regions where there is an excess of CpG dinucleotides compared with the rest of the genome. These are usually 300-3000 bp in length and often occur in promoter regions, where methylation can contribute to gene silencing.

- It is likely that CpG islands are maintained by purifying selection on their regulatory functions.
Suppose that we have sequenced a new genome and that we wish to identify the locations of the \(CpG\) islands. Analysis of previously sequenced genomes has revealed that dinucleotide frequencies differ greatly depending on whether a sequence lies within or outside of a \(CpG\) island. However, because the locations of these islands are unknown, we can use a HMM to model both the sequence of nucleotides and the hidden locations of the \(CpG\) islands (see Durbin et al., 1998 for details).

- The Markov chain \(X\) will take values in the set \(\{T^+, C^+, A^+, G^+, T^-, C^-, G^-, A^-\}\), where \(T^+\) will denote a thymine located within a \(CpG\) island and \(T^-\) will denote a thymine located outside of an island. These are hidden states, since we don’t know whether a given nucleotide resides in or outside of an island.

- The emitted or observed states are the nucleotides themselves. Thus the emission probabilities take a particularly simple form in this case:

\[
\begin{align*}
e(T|T^+) &= e(T|T^-) = 1 \\
e(C|C^+) &= e(C|C^-) = 1 \\
e(A|A^+) &= e(A|A^-) = 1 \\
e(G|G^+) &= e(G|G^-) = 1.
\end{align*}
\]

In particular, all other emission probabilities will be 0, e.g., \(e(A|T^+) = 0\).
The transition matrix of the hidden Markov chain will take the following form:

\[
P = \begin{pmatrix}
(1 - 4p)P^+ & p1 \\
q1 & (1 - 4q)P^-
\end{pmatrix},
\]

where the matrices (with states in order of \(TCAG\))

\[
P^+ = \begin{pmatrix}
0.182 & 0.355 & 0.079 & 0.384 \\
0.188 & 0.368 & 0.171 & 0.274 \\
0.120 & 0.274 & 0.180 & 0.426 \\
0.125 & 0.339 & 0.161 & 0.375
\end{pmatrix},
\]

and

\[
P^- = \begin{pmatrix}
0.292 & 0.239 & 0.177 & 0.292 \\
0.302 & 0.298 & 0.322 & 0.078 \\
0.210 & 0.205 & 0.300 & 0.285 \\
0.208 & 0.246 & 0.248 & 0.298
\end{pmatrix},
\]

were estimated from human sequence data and \(1\) is a \(4 \times 4\) matrix consisting of 1's. The parameters \(p\) and \(q\) are the probabilities of transitioning out of or into a \(CpG\) island, respectively, and satisfy \(1 > p > q > 0\).
Joint Probabilities of States and Observations

Suppose that we are given both a sequence of observations \( y_{0:T} \) and a (putative) sequence of states \( x_{0:T} \). Then the Markovian structure of the HMM makes it easy to calculate the joint probability of both the states and the observations:

\[
\mathbb{P}(x_{0:T}, y_{0:T}) = \mu(x_0) e(y_0|x_0) \prod_{t=1}^{T} (p_{x_{t-1},x_t} \cdot e(y_t|x_t))
\]

\[
= \left( \mu(x_0) \prod_{t=1}^{T} p_{x_{t-1},x_t} \right) \left( \prod_{t=0}^{T} e(y_t|x_t) \right).
\]

In other words, the joint probability is equal to the probability of the sample path traversed by the Markov chain multiplied by the probabilities of the states emitted by the chain conditional on that sample path.
Marginal Probabilities of Observations: The Forward Algorithm

Now suppose that we only know the sequence of observations $y_0:T$ and that we want to calculate the marginal probability of this sequence. Because the process $(Y_t : t \geq 0)$ usually is not Markovian, we cannot calculate this probability simply by multiplying transition probabilities, i.e.,

$$
P(y_0:T) \neq P(Y_0 = y_0) \prod_{t=1}^{T} P(Y_t = y_t | Y_{t-1} = y_{t-1}).$$

On the other hand, while it is true that

$$
P(y_0:T) = \sum_{x_0:T} P(x_0:T, y_0:T),$$

and we can calculate the joint probabilities appearing on the right-hand side, it usually will not be feasible to carry this out in practice due to the large number of terms ($|S|^{T+1}$) appearing in the sum.
Fortunately, it is possible to evaluate the probability of a sequence of evaluations without summing over all possible sample paths of the Markov chain by using a recursive algorithm known as the **forward algorithm**. To do so, we need to introduce a set of auxiliary quantities,

\[ f_{t,x} \equiv \mathbb{P}(y_{0:t}, X_t = x), \]

for \( t = 0, \cdots, T \) and \( x \in S \). Here \( f_{t,x} \) is the joint probability of the observations \( y_{0:t} \) up to time \( t \) along with the state \( x_t = x \). In view of the Markovian structure of the model, these probabilities satisfy the following recursion:

\[ f_{t,x} = e(y_t|x) \sum_{s \in S} f_{t-1,s} \cdot p_{s,x}, \]

since conditional on \( x_{t-1} = s \), the state \( x_t \) and the observation \( y_t \) are independent of the observations \( y_{0:t} \). The summation is necessary to average over all of the possible values of \( x_{t-1} \), but notice that this only involves \(|S|\) terms.
The forward algorithm consists of the following steps:

- **Initialization**: For each \( x \in S \), set
  \[
  f_{0,x} = e(y_0|x) \mu(x).
  \]

- **Recursion**: For \( t = 1, \cdots, T \) and each \( x \in S \), set
  \[
  f_{t,x} = e(y_t|x) \sum_{s \in S} f_{t-1,s} \cdot p_{s,x}.
  \]

- **Termination**: The probability of the observations is given by
  \[
  P(y_0:T) = \sum_{x \in S} f_{T,x}.
  \]

**Remark**: Execution of this algorithm requires order \( O(T|S|^2) \) operations: it is quadratic in \( |S| \) because we must calculate one quantity \( f_{t,x} \) for each \( x \in S \), but to do so requires evaluating a sum over \( S \). On the other hand, the algorithm scales linearly with \( T \), which allows it to be used to analyze very long sequences of observations.
Smoothing and Filtering: The Forward-Backward Algorithm

Given a sequence of observations $y_{0:T}$, we may wish to calculate the posterior distribution of the hidden state variable at time $t$, i.e.,

$$\mathbb{P}(X_t = x | y_{0:T}) = \frac{\mathbb{P}(y_{0:T}, X_t = x)}{\mathbb{P}(y_{0:T})}.$$

If $t = T$, then we say that we are filtering the data, i.e., we are using the sequence of observations to make inferences about the present state of the latent process. In this case, the joint distribution of the observations and the hidden state $X_T$ in the numerator is given by the quantity $f_{T,x}$ defined above and so the filtering problem can be solved using the forward algorithm:

$$\mathbb{P}(X_T = x | y_{0:T}) = \frac{f_{T,x}}{\mathbb{P}(y_{0:T})} = \frac{f_{T,x}}{\sum_{s \in S} f_{T,s}}.$$
On the other hand, if $t < T$, then we are using the data to make inferences about a past state of the latent process and we say that we are **smoothing** the data. In this case, the forward algorithm is no longer sufficient, but we can exploit the Markovian structure of the HMM to decompose the joint probability of $y_{0:T}$ and $X_t = x$ as follows:

\[
P(y_{0:T}, X_t = x) = P(y_{0:t}, X_t = x) \cdot P(y_{t+1:T} | X_t = x) = f_{t,x} \cdot b_{t,x},
\]

giving

\[
P(X_t = x | y_{0:T}) = \frac{f_{t,x} \cdot b_{t,x}}{P(y_{0:T})}.
\]

The first quantity, $f_{t,x}$, can be calculated with the forward algorithm, but the second quantity, here denoted, $b_{t,x}$ requires a new method. Fortunately, these probabilities also satisfy a simple recursion:

\[
b_{t,x} = \sum_{s \in S} p_{x,s} \cdot e(y_{t+1} | s) \cdot b_{t+1,s}.
\]
The algorithm used to recursively evaluate the probabilities $b_{t,x}$ is known as the **backward algorithm** because it starts in the present and works backward in time. It consists of the following steps:

- **Initialization:** For each $x \in S$, set $b_{T,x} = 1$.
- **Recursion:** For $t = T - 1, \cdots, 0$ and each $x \in S$, set
  \[
  b_{t,x} = \sum_{s \in S} p_{x,s} \cdot e(y_{t+1}|s) \cdot b_{t+1,s}.
  \]
- **Termination (if needed):** The probability of the observations is given by
  \[
  P(y_0:T) = \sum_{x \in S} \mu(x) e(y_0|x) b_{0,x}.
  \]

Like the forward algorithm, the backward algorithm requires order $O(T|S|^2)$ operations. Together, these can be used to calculate both $f_{t,x}$ and $b_{t,x}$ for any $t = 0, \cdots, T$, which can then be used to evaluate the posterior probabilities $P(X_t = x|y_0:T)$.
Decoding: The Viterbi Algorithm

Suppose that we are given a sequence of observations \( y_{0:T} \) and that we need to use this data to learn about the unknown sequence of states \( x_{0:T} \) visited by the Markov chain \( X \). Although the hidden states usually will not be uniquely determined by the observations, some sequences may be much more probable given the observations than others. Let us define the maximum a posteriori (MAP) estimate of \( x_{0:T} \) given \( y_{0:T} \) to be the sequence \( x^*_{0:T} \) that maximizes the probability of both the states and observations through time \( T \):

\[
x^*_{0:T} = \arg \max_{x_{0:T}} P(x_{0:T}, y_{0:T}).
\]

Although the MAP estimate can in principle be identified by brute force optimization, in practice this usually is not feasible since it requires \( |S|^{T+1} \) likelihood calculations.
Fortunately, there is a recursive algorithm which solves this problem without resorting to brute force maximization. This is known as the **Viterbi algorithm** and it requires that we calculate and save the following probabilities

\[
v_{t,x} \equiv \max_{x' \in S} P(x_0:t-1, y_0:t | X_t = x)
\]

for \( t = 0, \cdots, T \) and all \( x \in S \). Here \( v_{t,k} \) is the probability of the most probable sequence of states which ends in state \( x \) at time \( t \) and emits observations \( y_0:t \). The key observation is that these probabilities can be calculated recursively

\[
v_{t,x} = e(y_t | x) \cdot \max_{s \in S} \{ v_{t-1,s} \cdot p_{s,x} \}.
\]
Why does it work?

- First, observe that there is no general relationship between the MAP estimate for the data up to time $t$ and the data up to time $t-1$. In particular, it will usually not be the case that $x_{0:t-1}^*$ can be obtained from $x_{0:t}^*$ by dropping the final state variable.

- On the other hand, if we condition on the state of the process at time $t$, then because our process is a HMM, we know that all of the states and observations that occur prior to time $t$ are conditionally independent of all of the states and observations that occur after time $t$.

- This means that if we wish to find the most probable sequence of states that visits state $x$ at time $t$, we can do so by separately finding the most probable sequence of states during $0:t$ that ends in state $x$ and the most probable sequence of states during $t:T$ that begins in state $x$.

- We can then find the most probable sequence of states by maximizing these probabilities over all of the possible states $x$ at time $t$. 
The Viterbi algorithm can be implemented using the following procedure:

- **Initialization**: Set $v_{0,x} = \pi_x e(y_0|x)$ for $x \in S$.

- **Recursion**: For $t = 1, \cdots, T$ and $x \in S$, let
  \[
  v_{t,x} = e(y_t|x) \cdot \max_{s \in S} \left( v_{t-1,s} \cdot p_{s,x} \right) 
  \]
  \[
  Ptr_t(x) = \arg \max_{s \in S} \left( V_{t-1,s} \cdot p_{s,x} \right) 
  \]

- **Termination**: Let
  \[
  p^* = \max_{x \in S} v_{T,x} 
  \]
  \[
  x_T^* = \arg \max_{x \in S} v_{T,x}. 
  \]

- **Traceback**: For $t = T - 1, \cdots, 0$, let
  \[
  x_t^* = Ptr_{t+1}(x_{t+1}^*) 
  \]
Remarks:

- It can be shown that the sequence $x^* = (x_0^*, x_1^*, \cdots, x_T^*)$ constructed during the traceback is the MAP estimate for the observations $y_{0:T}$ and that $p^*$ is the probability of this sequence.

- If the maximization problems solved during execution of Viterbi’s algorithm do not have unique solutions, then in general the MAP estimate will not be unique. When this happens, we can find all of the MAP estimates by saving each of the maximizing values of $s$ to the relevant pointer variables (which need to be implemented as arrays). During the traceback, we will then have multiple choices at at least one of the steps and each will lead to a valid MAP estimate.

- Execution of the Viterbi algorithm requires $O(T \cdot |S|^2)$ computations.

- To avoid underflow, it is often necessary to work with logarithms of the probabilities, in which case we use the recursion:

$$\ln(v_{t,x}) = \ln(e(y_t|x)) + \max_{s \in S} \left\{ \ln(v_{t-1,s}) + \ln(p_{s,x}) \right\}.$$