Our goal is to use finite, discrete Markov chains to model the stochastic variation of a random variable. On Tuesday, we considered three examples of Markov models used in sequence analysis.

Examples:

1. Mutations at a single site in a DNA sequence. This Markov chain has four states: \( E_1 = A, E_2 = C, E_3 = G, E_4 = T \).
2. Local site dependence in all amino acid sequence. This Markov chain has twenty states: \( E_1 = \text{Ala}, E_2 = \text{Cys}, \ldots E_{20} = \text{Tyr} \).
3. Similarity score of prefixes of a pairwise alignment, with no gaps, of sequences \( \sigma \) and \( \tau \) of length \( L \). State \( E_j, -L \leq j \leq L \), corresponds to the difference between the number of matches and the number of mismatches.

Following the notation in Ewens and Grant:

- \( \varphi_j(t) \) is the probability that the chain is in state \( E_j \) at time \( t \). The vector \( \varphi(t) = (\varphi_1(t), \ldots \varphi_s(t)) \) is the state probability distribution at time \( t \).
- \( \varphi(0) \) is the initial state probability distribution.
- The transition probability \( P_{jk} = P(S_t = E_k | S_{t-1} = E_j) \), gives the probability that the chain will be in state \( k \) at time \( t \) given that it was in state \( E_j \) at the previous time step, \( t - 1 \). Note that \( \sum_j P_{jk} = 1 \). The chain must always be in some state.

The Markov property states that Markov chains are memoryless. In other words, the probability that the chain is in state \( E_j \) at time \( t \), depends only on the state at the previous time step, \( t - 1 \), not on history of states at \( t - 2, t - 3 \ldots \)

In this course, we will focus on discrete, finite, time-homogeneous Markov chains. These are models with a finite number of states, in which time (or space) is split into discrete steps. The assumption of discrete steps is somewhat artificial for the sequence evolution model in example 1, but quite natural for examples 2 and 3, because sequences of symbols are inherently discrete. Our models are time-homogeneous, because the transition matrix, \( P_{jk} \), does not change over time.

How does the state probability distribution change over time? If we know the state probability distribution at time \( t \), the distribution at the next time step is given by:

\[
\varphi_j(t + 1) = \sum_j \varphi_j(t) P_{jk}
\]  

(1)
or

\[ \varphi(t + 1) = \varphi(t)P \]  

(2)

in matrix notation.

Consider the example of a random walk with absorbing boundaries with five states similar to the gambling game we discussed in class on Tuesday. At each step, the model moves to the left or to the right with equal probability, resulting in the following transition probability matrix:

\begin{array}{cccccc}
0 & 1 & 2 & 3 & 4 \\
0 & 1 & 0 & 0 & 0 \\
1 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\
2 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\
3 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\
4 & 0 & 0 & 0 & 0 & 1 \\
\end{array}

Further, assume that the initial state distribution is \( \varphi(0) = (0, 0, 1, 0, 0) \). To obtain the state probability distribution after one time step, we apply Equation 1:

\[
\varphi_1(1) = \sum_{j=0}^{4} \varphi_j(0)P_{jk}
\]

\[
= 0 + 0 + 0 + 1 \cdot \frac{1}{2} + 0 + 0 + 0
\]

\[
= \frac{1}{2}
\]

Since the Markov chain is symmetrical, \( \varphi_3(1) \) is also equal to 1/2. It is not possible to reach state \( E_0 \) or state \( E_4 \) in a single step from state \( E_2 \). Nor is it possible to remain in state \( E_2 \) for two
consecutive time steps. Since state \( E_2 \) is the only state with non-zero probability at time \( t = 0 \), we obtain,

\[
\varphi(1) = (0, \frac{1}{2}, 0, \frac{1}{2}, 0).
\]

Now, following the same procedure, we calculate the probability distribution at time \( t = 2 \) from \( \varphi(1) \):

\[
\varphi_1(2) = \sum_{j=0}^{4} \varphi_j(1) P_{jk} = 0 + 0 + \frac{1}{2} \cdot \frac{1}{2} + 0 + 0 + 0 + 0 + 0 + 0 = \frac{1}{4}
\]

and

\[
\varphi_3(2) = \sum_{j=0}^{4} \varphi_2(1) P_{jk} = 0 + 0 + \frac{1}{2} \cdot \frac{1}{2} + 0 + \frac{1}{2} \cdot \frac{1}{2} + 0 + 0 = \frac{1}{2}
\]

As above, \( \varphi_4(1) = \varphi_0(1) \), by symmetry, yielding

\[
\varphi(2) = \left( \frac{1}{4}, 0, \frac{1}{2}, 0, \frac{1}{4} \right).
\]

**Steady state**

A state probability distribution, \( \varphi^* \), that satisfies the equation

\[
\varphi^* = \varphi^* P
\]

is called a stationary distribution. A key question for a given Markov chain is whether such a stationary distribution exists. Equation 3 is equivalent to a system of \( s \) equations in \( s \) unknowns. One way to determine the steady state distribution is to solve that system of equations. For example,
for the five state random walk with absorbing states, we would proceed as follows. The probability of being in state $E_0$ is

$$
\varphi_0 = \sum_{j=0}^{4} \varphi_j P_{j0}
$$

$$
= \varphi_0 P_{00} + \varphi_1 P_{10} + \varphi_2 P_{20} + \varphi_3 P_{30} + \varphi_4 P_{40}
$$

$$
= \varphi_0 + \frac{1}{2} \varphi_1,
$$

since $P_{20}$, $P_{30}$ and $P_{40}$ are all equal to zero. The other steady state probabilities are derived similarly, yielding

$$
\varphi_0 = \varphi_0 + \frac{1}{2} \varphi_1
$$

$$
\varphi_1 = \frac{1}{2} \varphi_2
$$

$$
\varphi_2 = \frac{1}{2} \varphi_1 + \frac{1}{2} \varphi_3
$$

$$
\varphi_3 = \frac{1}{2} \varphi_2
$$

$$
\varphi_4 = \varphi_4 + \frac{1}{2} \varphi_3.
$$

In addition, the probability that the system is in some state is unity, imposing an additional constraint:

$$
\varphi_0 + \varphi_1 + \varphi_2 + \varphi_3 + \varphi_4 = 1.
$$

The model has a stationary distribution if the above equations have a solution. The stationary state can also be obtained using matrix algebra, but that approach is beyond the scope of this course.

If we know the stationary state distribution, or have an educated guess, we can verify that it indeed satisfies Equation 3. For example, it is easy to show that the stationary state distribution for the random walk with absorbing boundaries is $(\frac{1}{2}, 0, 0, 0, \frac{1}{2})$. The equation $\varphi = (\frac{1}{2}, 0, 0, 0, \frac{1}{2})P$ yields

$$
\varphi_0 = \varphi_0 + \frac{1}{2} \varphi_1 = \frac{1}{2} + \frac{1}{2} \cdot 0 = \frac{1}{2}
$$

$$
\varphi_1 = \frac{1}{2} \varphi_2 = 0
$$

$$
\varphi_2 = \frac{1}{2} \varphi_1 + \frac{1}{2} \varphi_3 = 0 + 0 = 0
$$

$$
\varphi_3 = \frac{1}{2} \varphi_2 = 0
$$

$$
\varphi_4 = \varphi_4 + \frac{1}{2} \varphi_3 = \frac{1}{2} + \frac{1}{2} \cdot 0 = \frac{1}{2}
$$
or

$$\varphi = \left( \frac{1}{2}, 0, 0, 0, \frac{1}{2} \right).$$

**Higher order Markov chains**

Suppose we wish to know the state of the system after two time steps. From Equation 1, we obtain

$$\varphi_l(t + 1) = \sum_{j=0}^{s} \varphi_j(t) P_{ji} \quad (4)$$

and

$$\varphi_k(t + 2) = \sum_{l=0}^{s} \varphi_l(t + 1) P_{lk}. \quad (5)$$

Substituting the right hand side of Equation 4 for $\varphi_l(t + 1)$ in Equation 5 yields

$$\varphi_k(t + 2) = \sum_{l=0}^{s} \left( \sum_{j=0}^{s} \varphi_j(t) P_{jl} \right) P_{lk}. \quad (6)$$

We can reverse the order of the summations since the terms may be added in any order, yielding

$$\varphi_k(t + 2) = \sum_{j=0}^{s} \varphi_j(t) \left( \sum_{l=0}^{s} P_{jl} P_{lk} \right). \quad (6)$$

The term in the inner summation is simply the element in row $j$ and column $k$ of the matrix obtained by multiplying matrix $P$ by itself. Let $P^{(2)} = P^2 = P \times P$. Then

$$P_{jk}^{(2)} = \sum_{l=0}^{s} P_{jl} P_{lk},$$

so that Equation 6 may be rewritten as

$$\varphi_k(t + 2) = \sum_{j=0}^{s} \varphi_j(t) P_{jk}^{(2)}. \quad (6)$$

Matrix $P^{(2)}$ is the transition matrix of a 2nd order Markov chain that has the same states as the 1st order Markov chain described by $P$. However, a single time step in $P^{(2)}$ is equivalent to two time steps in $P$. Similarly, an $n^{th}$ Markov chain models change after $n$ time steps with a transition probability matrix

$$P^{(n)} = P^n = P \times P \ldots P.$$
As an example, we use this approach to investigate the periodicity of a 5-state random walk with reflecting boundaries. In the case of reflecting boundaries, the transition matrix, $P$, for the 1st order random walk is

$$
\begin{array}{c|cccc}
0 & 1 & 2 & 3 & 4 \\
\hline
0 & 0 & 1 & 0 & 0 \\
1 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\
2 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\
3 & 0 & 0 & \frac{1}{2} & 0 \\
4 & 0 & 0 & 0 & 1 \\
\end{array}
$$

Multiplying $P$ times itself yields the 2nd order transition matrix, $P^{(2)}$:

$$
\begin{array}{c|cccc}
0 & 1 & 2 & 3 & 4 \\
\hline
0 & \frac{1}{4} & 0 & \frac{1}{2} & 0 \\
1 & 0 & \frac{3}{4} & 0 & \frac{1}{2} \\
2 & \frac{1}{4} & 0 & \frac{1}{2} & 0 \\
3 & 0 & \frac{1}{2} & 0 & \frac{1}{4} \\
4 & 0 & 0 & \frac{1}{2} & 0 \\
\end{array}
$$

The initial state probability distribution and the state distribution after one time step are the same in both random walk models, namely $\varphi(0) = (0, 0, 1, 0, 0)$ and $\varphi(1) = (0, \frac{1}{2}, 0, \frac{1}{2}, 0)$. We calculate the state probability distribution at $t = 3$ by applying $P^{(2)}$:

$$
\begin{align*}
\varphi(3) &= \varphi(1) \cdot P^{(2)} \\
&= (0, \frac{1}{2}, 0, \frac{1}{2}, 0) \cdot P^{(2)} \\
&= (0, \frac{1}{2}, 0, \frac{1}{2}, 0)
\end{align*}
$$

This demonstrates that $\varphi(3) = \varphi(1)$ and, more generally, that the probability state distribution at all odd time intervals will be $(0, \frac{1}{2}, 0, \frac{1}{2}, 0)$.

Similarly, we can apply $P^{(2)}$ to calculate the probability distribution at even time intervals:

$$
\begin{align*}
\varphi(2) &= \varphi(0) \cdot P^{(2)} \\
&= (0, 0, 2, 0, 0) \cdot P^{(2)} \\
&= (\frac{1}{4}, 0, \frac{1}{2}, 0, \frac{1}{4})
\end{align*}
$$
and

\[ \varphi(4) = \varphi(2) \cdot P^{(2)} \]

\[ = \left( \frac{1}{4}, 0, \frac{1}{2}, 0, \frac{1}{4} \right) \cdot P^{(2)} \]

\[ = \left( \frac{1}{4}, 0, \frac{1}{2}, 0, \frac{1}{4} \right). \]

This shows that the state probability distribution is \( \left( \frac{1}{4}, 0, \frac{1}{2}, 0, \frac{1}{4} \right) \) at all positive even time steps. Thus, the random walk with reflecting boundaries is a periodic Markov chain with period 2.