Continuous-time Markov Processes

Beginning with these slides, we switch focus from discrete-time stochastic processes to those that are continuous-in-time. Although continuous-time processes are more complicated in some respects, they are needed in order to be able to effectively model two kinds of biological processes:

- processes with continuously-changing state, such as the diffusion of a particle through the cytoplasm of a cell;
- processes with discrete changes that occur at irregular times, such as births and deaths in a population that breeds at different rates throughout the year.

Before introducing the general theory, we will start by describing a particularly important example of a continuous-time Markov process called the Poisson process. These can be thought of as building blocks for the larger class of continuous-time Markov chains, but are also important models in their own right.
Example: The Poisson Process

Let \( \eta_1, \eta_2, \cdots \) be a sequence of independent exponentially-distributed random variables with rate \( \lambda \) and define a new sequence of random variables \( N = (N_t; t \geq 0) \), with continuous time parameter \( t \), by

\[
N_t = \sup\{ n \geq 0 : \eta_1 + \cdots + \eta_n < t \}.
\]

The process \( N \) defined in this manner is said to be a Poisson process with rate \( \lambda \) and is an example of a counting process. It is often interpreted in the following manner. Think of \( \eta_i \) as the random waiting time between successive events and let

\[
T_n = \eta_1 + \cdots + \eta_n
\]

be the time of the \( n \)'th event. Then \( N_t \) is the number of events that have occurred up to time \( t \).
To calculate the distribution of $N_t$, we need to recall that the distribution of a sum of $n$ i.i.d. exponential random variables with parameter $\lambda$ is the Gamma distribution with parameters $n$ and $\lambda$. Then, since

$$\{N_t = n\} = \{T_n \leq t < T_n + \eta_{n+1}\}$$

it follows that

$$\mathbb{P}(N_t = n) = \mathbb{P}(T_n \leq t < T_n + \eta_{n+1})$$

$$= \int_0^t \mathbb{P}(T_n = s)\mathbb{P}(\eta_{n+1} > t - s)ds$$

$$= \int_0^t \frac{\lambda^n}{\Gamma(n)} s^{n-1} e^{-\lambda s} e^{-\lambda(t - s)} ds$$

$$= e^{-\lambda t} \frac{\lambda^n}{(n - 1)!} \int_0^t s^{n-1} ds$$

$$= e^{-\lambda t} \frac{(\lambda t)^n}{n!}.$$

Since this holds for every $n \geq 0$, it follows that $N_t$ is Poisson-distributed with parameter $\lambda t$. Of course, this is also the source of the name of this process.
With some additional work, it can be shown that the process $N$ also has the following properties:

1. $N$ has independent increments: for all $0 \leq t_0 < t_1 < \cdots < t_k$, the random variables $N_{t_0}, N_{t_1} - N_{t_0}, N_{t_2} - N_{t_1}, \cdots, N_{t_k} - N_{t_{k-1}}$ are independent.

2. For each $t > 0$,

$$
\mathbb{P}(N_{t+\delta t} = n + 1 | N_t = n) = \lambda \delta t + o(\delta t)
$$

$$
\mathbb{P}(N_{t+\delta t} = n | N_t = n) = 1 - \lambda \delta t + o(\delta t)
$$

$$
\mathbb{P}(N_{t+\delta t} = i | N_t = n) = o(\delta t), \text{ for every } i \neq n, n + 1.
$$

Both of these properties are related to the fact that the Poisson process is itself a Markov process. The first property asserts that the numbers of events occurring in disjoint time intervals are independent; in fact, this is stronger than the Markov property. The second property asserts that the probability of having more than one jump in any short time interval is negligibly small. It also asserts that the rate of jumps from $n$ to $n + 1$ is equal to $\lambda$. 
We now turn to the general definitions.

**Definition**

A stochastic process $X = (X_t; t \geq 0)$ with values in a set $E$ is said to be a continuous-time Markov process if for every sequence of times $0 \leq t_1 < \cdots < t_n < t_{n+1}$ and every set of values $x_1, \cdots, x_n \in E$, we have

$$\mathbb{P}(X_{t_{n+1}} \in A | X_{t_1} = x_1, \cdots, X_{t_n} = x_n) = \mathbb{P}(X_{t_{n+1}} \in A | X_{t_n} = x_n),$$

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

$$p(s, t; x, A) = \mathbb{P}(X_t \in A | X_s = x), \quad t \geq s \geq 0$$

is called the transition function of $X$. If this function depends on $s$ and $t$ only through the difference $t - s$, i.e., if we can write

$$p(t - s, x, A) = \mathbb{P}(X_t \in A | X_s = x)$$

for every $t \geq s \geq 0$, then we say that $X$ is time-homogeneous.
Sample Paths

Remarks: A continuous-time stochastic process $X = (X_t : t \geq 0)$ can be thought of in two different ways.

- On the one hand, $X$ is simply a collection of random variables defined on the same probability space, $\Omega$.
- On the other hand, we can also think of $X$ as a path- or function-valued random variable. In other words, given an outcome $\omega \in \Omega$, we will view $X(\omega)$ as a function from $[0, \infty)$ into $E$ defined by

$$X(\omega)(t) \equiv X_t(\omega) \in E.$$  

The path traced out by $X(\omega)$ as $t$ varies from 0 to $\infty$ is said to be a sample path of $X$. 
In these notes, we will mainly be concerned with Markov processes that take values in countable state spaces. As in our discussion of discrete-time Markov chains, we will assume without loss of generality that the state space is either $E = \{1, 2, \cdots, n\}$ or $E = \{1, 2, \cdots\}$.

**Definition**

A stochastic process $X = (X_t; t \geq 0)$ with values in a countable set $E = \{1, 2, \cdots\}$ is said to be a **time-homogeneous continuous-time Markov chain (CTMC)** if for every sequence of times $0 \leq t_1 < \cdots < t_n < t_{n+1}$ and every set of values $x_1, \cdots, x_{n+1} \in E$, there is a function

$$p : [0, \infty) \times E \times E \to [0, 1]$$

such that

$$\mathbb{P}(X_{t_{n+1}} = x_{n+1}|X_{t_1} = x_1, \cdots, X_{t_n} = x_n) = p(t_{n+1} - t_n, x_n, x_{n+1}).$$
An important difference between the treatment of discrete-time and continuous-time Markov chains is that in the latter case there is no one canonical transition matrix that is used to characterize the entire process. Instead, if \( X = (X_t : t \geq 0) \) is a continuous-time Markov chain, we can define an entire family of transition matrices indexed by time. Specifically, for each \( t \geq 0 \) and each pair of elements \( i, j \in E \), let

\[
p_{ij}(t) \equiv p(t, i, j) = \mathbb{P}(X_t = j | X_0 = i)
\]

and let \( P(t) \) be the matrix with entries \( p_{ij}(t) \). In particular, if \( E = \{1, 2, \cdots, n\} \) is finite, then \( P(t) \) is the \( n \times n \) matrix

\[
P(t) = \begin{pmatrix}
  p_{11}(t) & p_{12}(t) & \cdots & p_{1n}(t) \\
p_{21}(t) & p_{22}(t) & \cdots & p_{2n}(t) \\
  \vdots & \vdots & & \vdots \\
p_{n1}(t) & p_{n2}(t) & \cdots & p_{nn}(t)
\end{pmatrix}.
\]
It can be shown that the family of transition matrices \((P(t) : t \geq 0)\) satisfies the following properties:

- For each \(t \geq 0\), \(P(t)\) is a stochastic matrix:
  \[
  \sum_{j \in E} p_{ij}(t) = 1.
  \]

- \(P(0)\) is the identity matrix: \(p_{ij}(0) = \delta_{ij}\).

- For every \(s, t \geq 0\), the matrices \(P(s)\) and \(P(t)\) commute and
  \[
  P(t + s) = P(t)P(s).
  \]

The third property is called the **semigroup property** and the family of matrices is said to be a **transition semigroup**. When written in terms of coordinates, this property is

\[
p_{ij}(s + t) = \sum_{k \in E} p_{ik}(s)p_{kj}(t)
\]

which we recognize as the continuous-time analogue of the Chapman-Kolmogorov equations.
One consequence of the semigroup property is that for any $t \geq 0$ and any $n \geq 1$, we have

$$P(t) = P(t/n)^n.$$ 

In fact, provided that the matrices $P(t)$ depend continuously on $t$, it can be shown that there exists a unique matrix $Q$ such that for every $t \geq 0$

$$P(t) = e^{Qt} = \sum_{n=0}^{\infty} \frac{1}{n!} t^n Q^n.$$ 

This matrix is called the rate matrix (or, alternatively, the generator matrix, infinitesimal matrix, or $Q$-matrix) of the continuous-time Markov chain $X$. Rate matrices play a leading role in the description and analysis of continuous-time Markov chain and have a special structure which is described in the next theorem.
**Theorem**

Let $Q = (q_{ij})$ be the rate matrix of a continuous-time Markov chain $X$ with transition semigroup $(P(t) : t \geq 0)$. Then all of the row-sums of $Q$ are equal to 0 and all of the off-diagonal elements are non-negative:

$$ q_{ij} \geq 0 \quad \text{if } j \neq i $$

$$ q_{ii} = -\sum_{j \neq i} q_{ij}, $$

Furthermore, each transition probability $p_{ij}(t)$ is a differentiable function of $t$ and

$$ q_{ij} = p_{ij}'(0) = \lim_{t \to 0} \frac{p_{ij}(t) - p_{ij}(0)}{t}. $$

**Remark:** In matrix notation, these identities can be written as

$$ Q = P'(0) = \lim_{t \to 0} \frac{1}{t} (P(t) - I). $$
This theorem explains why $Q$ is called the rate matrix of the process $X$. Provided that $t > 0$ is sufficiently small,

$$P(X_t = j|X_0 = i) = \begin{cases} q_{ij}t + o(t) & \text{if } j \neq i \\ 1 + q_{ii}t + o(t) & \text{if } j = i, \end{cases}$$

which shows that the probability that the process moves from its current state $i$ to another state $j$ during a short period of time $t$ is approximately proportional to the amount of time elapsed. In other words, $q_{ij}$ is the rate at which transitions occur from state $i$ to state $j$, while $|q_{ii}|$ is the total rate at which the process jumps out of state $i$.

**Example:** The rate matrix of the Poisson process $N$ with rate $\lambda$ has entries

$$q_{ij} = \begin{cases} \lambda & \text{if } j = i + 1 \\ -\lambda & \text{if } j = i \\ 0 & \text{otherwise}. \end{cases}$$
Q-Matrices and Continuous-Time Markov Chains

Let us say that a square matrix $Q = (q_{ij})$ is a Q-matrix if all of the row sums are equal to 0 and if all of the off-diagonal elements are non-negative. Given any Q-matrix with bounded entries, we can always construct a continuous-time Markov chain $X = (X_t : t \geq 0)$ which has this matrix as its rate matrix. This construction is described below and requires the following objects:

- A probability distribution $\nu$ on $E$ which will be the initial distribution of the chain.
- A collection of independent exponentially-distributed random variables $\eta_{ij}^{(n)}; n \geq 1; i, j \in E, j \neq i$, where $\eta_{ij}^{(n)}$ has rate parameter $q_{ij}$. If $q_{ij} = 0$, we simply set $\eta_{ij}^{(n)} = \infty$. 
The process $X$ is constructed recursively according to the following procedure:

- Sample a state $i_0 \in E$ with probability $\nu_{i_0}$ and set $X_0 = i_0$; this is the initial state of the chain.
- Conditional on $X_0 = i_0$, let
  \[
  \tau_1 \equiv \min \left\{ \eta^{(1)}_{i_0j} : j \neq i_0 \right\} \quad \text{and} \quad i_1 \equiv \arg \min_j \left\{ \eta^{(1)}_{i_0j} : j \neq i_0 \right\} .
  \]
- Let $J_1 = \tau_1$ and define $X$ on the interval $[0, J_1]$ by setting
  \[
  X_t = \begin{cases} 
  i_0 & \text{if } t \in [0, J_1) \\
  i_1 & \text{if } t = J_1.
  \end{cases}
  \]

**Interpretation:** The process occupies state $i_0$ up until the time of the first jump, $J_1$, at which point it jumps to state $i_1$. 
In general,

- Conditional on $X_{J_n} = i_n$, let
  
  $$\tau_{n+1} \equiv \min \left\{ \eta_{inj}^{(n+1)} : j \neq i_n \right\}$$

  $$i_{n+1} \equiv \arg \min_j \left\{ \eta_{inj}^{(n+1)} : j \neq i_n \right\}.$$ 

- Let $J_{n+1} = J_n + \tau_{n+1}$, and define $X$ on the interval $[J_n, J_{n+1}]$ by setting

  $$X_t = \begin{cases} 
  i_n & \text{if } t \in [J_n, J_{n+1}) \\
  i_{n+1} & \text{if } t = J_{n+1}.
  \end{cases}$$

**Interpretation:** The process occupies state $i_n$ between times $J_n$ and $J_{n+1} = J_n + \tau_{n+1}$ and then jumps to state $i_{n+1}$. 
Under this construction, the sample paths of $X$ are piecewise constant, with jumps at the times $J_1, J_2, \cdots$. For this reason, the random times $J_1, J_2, \cdots$ are known as the **jump times** of $X$, while the times $\tau_1, \tau_2, \cdots$ are called the **interarrival times** or **holding times** of this process. Both the holding times and the states visited but the Markov chain depend on the rate matrix through the relationship:

$$
\tau_{n+1} = \min \left\{ \eta_{i_nj}^{(n+1)} : j \neq i_n \right\}, \\
i_{n+1} = \arg \min_j \left\{ \eta_{i_nj}^{(n+1)} : j \neq i_n \right\}.
$$

This has the following interpretation. Given that the process enters state $i_n$ following the $n$'th jump, we assign an independent exponential random clock $\eta_{i_nj}^{(n+1)}$ with rate $q_{i_nj}$ to each alternative state $j \neq i_n$. Then, both the holding time and the next state visited by the chain are determined by the clock that rings first. Notice that the larger the rate $q_{ij}$ is, the more likely it is that the process will jump to state $j$.

**Remark:** It can be shown that the process $X$ constructed in this fashion has rate matrix $Q$. 
The Jukes-Cantor Model and Divergence Time Estimates

The Jukes-Cantor model (JC69) is one of the simplest models of the neutral substitution process at a single site in a DNA molecule. It is named after Thomas Jukes and Charles Cantor, who introduced it in 1969. To describe this process, we will identify the state space \( \{T, C, A, G\} \) with the set \( E = \{1, 2, 3, 4\} \). JC69 assumes that all single nucleotide mutations occur at the same rate \( \mu \), so that the rate matrix is just

\[
Q = \begin{pmatrix}
-3\mu & \mu & \mu & \mu \\
\mu & -3\mu & \mu & \mu \\
\mu & \mu & -3\mu & \mu \\
\mu & \mu & \mu & -3\mu
\end{pmatrix},
\]

while the transition matrices are given by

\[
P(t) = e^{Qt} = \frac{1}{4} \begin{pmatrix}
1 + 3e^{-4\mu t} & 1 - e^{-4\mu t} & 1 - e^{-4\mu t} & 1 - e^{-4\mu t} \\
1 - e^{-4\mu t} & 1 + 3e^{-4\mu t} & 1 - e^{-4\mu t} & 1 - e^{-4\mu t} \\
1 - e^{-4\mu t} & 1 - e^{-4\mu t} & 1 + 3e^{-4\mu t} & 1 - e^{-4\mu t} \\
1 - e^{-4\mu t} & 1 - e^{-4\mu t} & 1 - e^{-4\mu t} & 1 + 3e^{-4\mu t}
\end{pmatrix}.
\]
One of the uses for substitution models like the JC69 model is in estimating the
divergence time between two individuals (possibly belonging to different species).

Suppose that we compare the genomes of two individuals that last shared a common
ancestor \( t \) units of time ago. Since the total time elapsed along the genealogy is \( 2t \) and
mutations could occur along either of the branches leading to these individuals, under
the JC69 model the probabilities that both individuals have the same nucleotide in a
homologous site or different nucleotides in that site are:

\[
\begin{align*}
\mathbb{P}(\text{same nucleotide}) &= \frac{1}{4} \left( 1 + 3e^{-8\mu t} \right) \\
\mathbb{P}(\text{different nucleotides}) &= \frac{3}{4} \left( 1 - e^{-8\mu t} \right).
\end{align*}
\]
Now suppose that we compare $L$ homologous sites between these two individuals and we find that they differ at $d$ sites and are identical at $L - d$ sites. If we assume that each site evolves at the same rate according to the Jukes-Cantor model and that the substitution processes at the different sites are independent, then the likelihood function for the divergence time $t_{\text{div}} = t$ is

$$L(t; d) = \left( \frac{3 (1 - e^{-8\mu t})}{4} \right)^d \left( \frac{1 + 3e^{-8\mu t}}{4} \right)^{L-d},$$

while the log-likelihood function is

$$l(t; d) = C + d \log \left( 1 - e^{-8\mu t} \right) + (L - d) \log \left( 1 + 3e^{-8\mu t} \right),$$

where $C$ is a constant that does not depend on $t$. 
The maximum likelihood estimate of $t_{\text{div}}$ can be found by differentiating the log-likelihood function with respect to $t$ and setting this equal to 0. This gives the equation

$$0 = d \frac{8e^{-8\mu t}}{1 - e^{-8\mu t}} + (L - d) \frac{-24e^{-8\mu t}}{1 + 3e^{-8\mu t}},$$

which can be solved explicitly. (To do so, first let $x = e^{-8\mu t}$ and solve for $x$, and then solve for $t$.) After some algebra, we find that the maximum likelihood estimate of the divergence time is equal to

$$\hat{t}_{\text{div}} = \frac{-1}{8\mu} \log \left( 1 - \frac{4d}{3L} \right),$$

provided that $d/L < 3/4$. 
Notice that when $d/L \ll 1$, the estimated divergence time is approximately linear in the proportion of sites that differ between the individuals:

$$\hat{t}_{\text{div}} \approx \frac{1}{2} \frac{1}{3\mu} \frac{d}{L},$$

However, this approximation breaks down as $d/L$ increases and in fact

$$\lim_{d/L \to 3/4} \hat{t}_{\text{div}} = \infty.$$  

This behavior is due to a phenomenon called **saturation**, which occurs when the divergence time is large enough that many sites undergo multiple substitutions, some of which will go unobserved because of reverse mutations that restore the ancestral nucleotide. For example, a $T$ may mutate to an $A$ which can then mutate back to $T$ and neither substitution will be discernible in the data. The level $d/L = 3/4$ is special because in this model two unrelated genomes ($t_{\text{div}} = \infty$) will share the same nucleotide at approximately a quarter of their sites just by chance.
We can use this result to obtain at least a crude estimate of the divergence time between humans and chimpanzees. On average, orthologous sequences found in both genomes differ at approximately 1.1% of their sites, giving 

\[
\frac{d}{L} \approx 0.011.
\]

Estimating the genome-wide average nucleotide mutation rate is trickier, but genome sequencing studies of several human parent-offspring trios suggest that this rate is approximately \(1.1 \times 10^{-8}\) mutations per site per generation. If we assume an ancestral generation time of 25 years, then taking into account that \(3\mu\) is the total mutation rate per site in the JC69 model, we obtain 

\[
\mu \approx \frac{1}{3} \frac{1.1 \times 10^{-8}}{25} = 1.5 \times 10^{-10} \text{ mutations/site \times yr}.
\]

Substituting these results into the formula for the maximum likelihood estimate gives 

\[
\hat{t}_{\text{div}}(\text{human, chimp}) \approx 11 \text{ million years}.
\]
The Kolmogorov Forward and Backward Equations

Suppose that $X$ is a continuous-time Markov chain with rate matrix $Q$ and transition functions $p_{ij}(t), i, j \in E$. If we split the time interval $[0, t + h]$ into the sub-intervals $[0, t]$ and $[t, h]$ and apply the Chapman-Kolmogorov equations, then for small values of $h > 0$, we have

\[
p_{ij}(t + h) = \sum_{k \in E} p_{ik}(t)p_{kj}(h)
= \sum_{k \in E} p_{ik}(t)\{\delta_{kj} + q_{kj}h + o(h)\}
= p_{ij}(t) + \sum_{k \in E} p_{ik}(t)\{q_{kj}h + o(h)\}.
\]

Notation: $\delta_{kj}$ is the Kronecker delta, which is equal to 1 when $k = j$ and 0 otherwise.
By rearranging terms and taking the limit as \( h \downarrow 0 \), we find that the transition probabilities \( p_{ij}(t), i, j \in E \) satisfy the following system of linear ODE’s,

\[
p'_{ij}(t) = \lim_{h \to 0^+} \frac{p_{ij}(t + h) - p_{ij}(t)}{h} = \sum_{k \in E} p_{ik}(t) \cdot q_{kj}
\]

which are known variously as the Kolmogorov forward equations, the master equations, or the Fokker-Planck equations.

**Remark:** These are called the forward equations because they ‘peek’ forward in time from the intermediate states \( k \) to the present state \( j \).
Similarly, if we split the time interval \([0, t + h]\) into the sub-intervals \([0, h]\) and \([h, t + h]\) and apply the Chapman-Kolmogorov equations, then as long as \(h > 0\) is small, we have

\[
p_{ij}(t + h) = \sum_{k \in E} p_{ik}(h)p_{kj}(t) = \sum_{k \in E} (\delta_{ik} + q_{ik}h + o(h)) p_{kj}(t) = p_{ij}(t) + \sum_{k \in E} (q_{ik}h + o(h)) p_{kj}(t).
\]

In this case, taking the limit as \(h \downarrow 0\) gives

\[
p'_{ij}(t) = \sum_{k \in E} q_{ik}p_{kj}(t)
\]

which are known as the \textbf{Kolmogorov backward equations} because they ‘peek’ backwards in time from the intermediate states \(k\) to the initial state \(i\).
The forward and backward equations can also be expressed in matrix notation as:

\[
\dot{P}(t) = P(t)Q \quad \text{(Forward equations)}
\]
\[
\dot{P}(t) = QP(t) \quad \text{(Backward equations)}
\]

subject to the initial condition \( P(0) = I \). Of course, both systems can be solved by matrix exponentiation, since we already know that \( P(t) = \exp(Qt) \), but sometimes it is possible to solve one or the other systems of equations without having to evaluate the matrix exponential. In particular, these equations are valid even when \( E \) is countably infinite.
Example: The marginal distributions of a Poisson process $N = (N_t : t \geq 0)$ can be found by solving the forward equation. Recall that the transition rates for this process are

$$q_{ij} = \begin{cases} 
\lambda & \text{if } j = i + 1 \\
-\lambda & \text{if } j = i \\
0 & \text{otherwise.}
\end{cases}$$

By convention, we set $N_0 = 0$ (i.e., no events have occurred at time 0), so that the marginal distribution of $N_t$ is given by the transition probabilities $p_{0n}(t)$. The forward equation for this process takes the following form:

$$\dot{p}_{00}(t) = -\lambda p_{00}(t)$$
$$\dot{p}_{0n}(t) = \lambda p_{0,n-1}(t) - \lambda p_{0n}(t), \quad n > 0,$$

with initial condition

$$p_{0n}(0) = \delta_{0n}.$$
The first equation only involves $p_{00}(t)$ and can be solved immediately to give

$$p_{00}(t) = e^{-\lambda t}.$$  

In contrast, the subsequent equations involve both $p_{0n}(t)$ and $p_{0,n-1}(t)$ and will lead to a recursion. Since the equation satisfied by $\dot{p}_{0n}(t)$ is a first-order linear differential equation, it can be solved by first multiplying both sides by the integrating factor $e^{\lambda t}$ and then rearranging so that we obtain

$$e^{\lambda t} \dot{p}_{0n}(t) + \lambda e^{\lambda t} p_{0n}(t) = \lambda e^{\lambda t} p_{0,n-1}(t).$$

Notice that the left-hand side is now an exact differential, i.e.,

$$\left( e^{\lambda t} p_{0n}(t) \right)' = \lambda e^{\lambda t} p_{0,n-1}(t),$$

which can be integrated (using the fact that $p_{0n}(0) = 0$ when $n > 0$) to give

$$p_{0n}(t) = e^{-\lambda t} \lambda \int_{0}^{t} e^{\lambda s} p_{0,n-1}(s) ds.$$
This last equation allows us to calculate $p_{0n}(t)$ once we know $p_{0,n-1}(t)$. However, since we already know that $p_{00}(t) = e^{-\lambda t}$, we can solve recursively for all of the probabilities $p_{0n}(t)$ by calculating these integrals. For example,

$$p_{01}(t) = e^{-\lambda t} \lambda \int_0^t e^{\lambda s} e^{-\lambda s} ds = e^{-\lambda t} \lambda t.$$  

Similarly,

$$p_{02}(t) = e^{-\lambda t} \lambda \int_0^t e^{\lambda s} e^{-\lambda s} \lambda ds = e^{-\lambda t} \frac{(\lambda t)^2}{2},$$  

and

$$p_{03}(t) = e^{-\lambda t} \lambda \int_0^t e^{\lambda s} e^{-\lambda s} \frac{\lambda^2 s^2}{2} ds = e^{-\lambda t} \frac{(\lambda t)^3}{3!}.$$
This leads us to conjecture that

\[ p_{0n}(t) = e^{-\lambda t} \frac{(\lambda t)^n}{n!}, \]

which can be confirmed by induction since

\[ p_{0,n+1}(t) = e^{-\lambda t} \lambda \int_0^t e^{\lambda s} e^{-\lambda s} \frac{(\lambda s)^n}{n!} \, ds = e^{-\lambda t} \frac{(\lambda t)^{n+1}}{(n + 1)!}. \]

Thus, as shown previously using a very different approach, we see that the marginal distribution of \( N_t \) is Poisson with parameter \( \lambda t \).
In the previous example, we used the Kolmogorov forward equation to solve for the marginal distribution of a Poisson process under the assumption that the initial state is 0. The next theorem generalizes this approach to arbitrary continuous-time Markov chains with arbitrary initial distributions.

**Theorem**

Suppose that $X$ is a continuous-time Markov chain with rate matrix $Q$ and initial distribution $\nu = (\nu_1, \nu_2, \cdots)$ on the state space $E$. Then the marginal distributions $p_j(t) = \mathbb{P}_\nu(X = j)$ solve the following initial value problem:

$$
\dot{p}_j(t) = \sum_{i \in E} q_{ij}p_i(t),
$$

with $p_j(0) = \nu_j$.

**Proof:** We first observe that the marginal distributions can be written as mixtures of the transition probabilities with weights provided by the initial distribution:

$$
p_j(t) = \mathbb{P}(X_t = j \mid X_0 \sim \nu) = \sum_{k \in E} \nu_k \mathbb{P}(X_t = j \mid X_0 = k) = \sum_{k \in E} \nu_k p_{kj}(t).
$$
Then, upon differentiating with respect to $t$ and using the forward equation, we have:

\[
\dot{p}_j(t) = \sum_{k \in E} \nu_k \dot{p}_{kj}(t) \\
= \sum_{k \in E} \nu_k \sum_{i \in E} p_{ki}(t) q_{ij} \\
= \sum_{i \in E} q_{ij} \sum_{k \in E} \nu_k p_{ki}(t) \\
= \sum_{i \in E} q_{ij} p_i(t).
\]
These equations can also be interpreted as a **conservation law** for probability mass. Indeed, if the $j$’th equation is rewritten in the form

$$
\dot{p}_j(t) = \sum_{i \neq j} q_{ij} p_i(t) - |q_{jj}| p_j(t),
$$

then we see that the rate of change in the probability mass occupying state $j$ is equal to the difference between the rate at which probability mass enters this state minus the rate at which probability mass exits this state. This holds because the total probability mass must be 1 at all times, giving

$$
0 = \left( \sum_{j \in E} p_j(t) \right)' = \sum_{j \in E} \dot{p}_j(t),
$$

so that probability mass can only be shifted around the state space, not created or destroyed.
Example: Suppose that $X$ is a Markov chain with values in a set containing just two states $\{1, 2\}$ and that the rate matrix of $X$ is

$$Q = \begin{pmatrix} -a & a \\ b & -b \end{pmatrix}.$$

In other words, $X$ simply alternates between these two states, possibly at different rates. If $p(t) = (p_1(t), p_2(t))$ is the marginal distribution of $X(t)$, then

$$\begin{align*}
\dot{p}_1(t) &= bp_2(t) - ap_1(t) \\
\dot{p}_2(t) &= ap_1(t) - bp_2(t)
\end{align*}$$

subject to the initial conditions $p_1(0) = \nu_1$ and $p_2(0) = \nu_2$. However, as we also know that $p_1(t) + p_2(t) = 1$, we can reduce this system to a single equation

$$\dot{p}_1(t) = b - (a + b)p_1(t)$$

which has solution

$$p_1(t) = \frac{b}{a + b} + \frac{\nu_1 a - \nu_2 b}{a + b} e^{-(a+b)t}.$$
Our next result tells us how to calculate the expected value of a function of a Markov chain, $\mathbb{E}[f(X_t)]$.

**Theorem**

Suppose that $X$ is a continuous-time Markov chain with rate matrix $Q$ on the state space $E$. Let $f : E \rightarrow \mathbb{R}$ be a bounded function from $E$ into $\mathbb{R}$ and for each $i \in E$ define the function $u : [0, \infty) \rightarrow \mathbb{R}$ by

$$ u_i(t) = \mathbb{E}[f(X_t)|X_0 = i]. $$

Then the functions $u_i(t), i \in E$ solve the following initial value problem

$$ \dot{u}_i(t) = \sum_{j \in E} q_{ij}(u_j(t) - u_i(t)) $$

with $u_i(0) = f(i)$. 
Proof: For each $i \in E$, the function $u_i(t)$ can be written as

$$u_i(t) = \sum_{k \in E} p_{ik}(t)f(k).$$

Then, upon differentiating with respect to $t$ and using the backward equation, we have:

$$\dot{u}_i(t) = \sum_{k \in E} \dot{p}_{ik}(t)f(k) = \sum_{k \in E} \sum_{j \in E} q_{ij}p_{jk}(t)f(k)$$

$$= \sum_{j \in E} q_{ij} \sum_{k \in E} p_{jk}(t)f(k)$$

$$= \sum_{j \in E} q_{ij} u_j(t)$$

$$= \sum_{j \neq i} q_{ij} u_j(t) - |q_{ii}| u_i(t)$$

$$= \sum_{j \neq i} q_{ij} u_j(t) - \sum_{j \neq i} q_{ij} u_i(t)$$

$$= \sum_{j \in E} q_{ij}(u_j(t) - u_i(t)).$$
Example: Continuous-time Moran Model

Although we previously introduced a discrete-time version of the Moran model, it is somewhat more natural to model a continuously-reproducing population in continuous time. Indeed, Moran’s original model was formulated as a continuous-time Markov chain. We make the following assumptions:

- The population consists of $N$ haploid individuals of genotypes $A_1$ or $A_2$.
- Each individual reproduces continuously, at rate 1, giving birth to a single offspring inheriting the parental genotype. Also, concurrent with any birth, one of the $N$ individuals in the population is chosen uniformly at random and dies.
- $A_1$ and $A_2$ are selectively neutral, so that births and deaths do not depend on the genetic composition of the population.
- Each $A_1$-type individual mutates at rate $\mu_{12}$ to $A_2$, while each $A_2$-type individual mutates at rate $\mu_{21}$ to $A_1$. 

Jay Taylor (ASU) 
APM 541 
Fall 2013
Let $Y_t$ denote the number of $A_1$-type individuals alive at time $t$ and suppose that $Y_t = i$. Then $Y_t$ can change either through a birth-death event or through a mutation event, and either event will result in $i$ increasing or decreasing by 1. Since no other transitions are possible, the rate matrix $Q = (q_{ij})$ is tridiagonal with

\[
q_{i,i+1} = N \left( \frac{i}{N} \right) \left( \frac{N-i}{N} \right) + (N - i) \mu_{21} \\
= Np_i(1 - p_i) + N(1 - p_i)\mu_{21}
\]

\[
q_{i,i-1} = N \left( \frac{i}{N} \right) \left( \frac{N-i}{N} \right) + i\mu_{12} \\
= Np_i(1 - p_i) + Np_i\mu_{12}
\]

\[
q_{ii} = -2Np_i(1 - p_i) - Np_i\mu_{12} - N(1 - p_i)\mu_{21}
\]

\[
q_{ij} = 0 \quad \text{if} \ |j - i| > 1,
\]

where $p_i = i/N$ is the frequency of $A_1$ in the population. As with the discrete-time model, it is customary to formulate this as a frequency-valued process $X = (X_t : t \geq 0)$ by setting $X_t = Y_t/N$. Notice that $X$ and $Y$ have the same rate matrix.
Let us calculate the mean frequency of $A_1$ at time $t$ 

$$u(p; t) \equiv \mathbb{E}[X_t : X_0 = p],$$

first under the assumption that both mutation rates are equal to 0. By the previous theorem we know that $u(p; t)$ satisfies the following system of differential equations:

$$\dot{u}(p; t) = \sum_{x \in E} q_{p,x} \{u(x; t) - u(p; t)\}$$

$$= Np(1 - p)\{u(p + 1/N; t) - u(p; t)\} + Np(1 - p)\{u(p - 1/N; t) - u(p; t)\}$$

subject to the initial condition $u(p; 0) = p \in \{0, 1/N, \cdots, 1\}$. I claim that the unique solution to this initial value problem is

$$u(p; t) \equiv p,$$

for all $t \geq 0$. In other words, as in the discrete-time model, the expected allele frequency is constant in a population subject only to genetic drift.
To verify this claim, observe that if \( u(p; t) = p \) for all \( t \geq 0 \), then \( \dot{u}(p; t) = 0 \). Furthermore,

\[
Np(1 - p) \{ u(p + 1/N; t) - u(p; t) \} = p(1 - p)
\]

\[
Np(1 - p) \{ u(p - 1/N; t) - u(p; t) \} = -p(1 - p)
\]

and so these two quantities sum to 0 for all \( p \) and all \( t \geq 0 \). This shows that \( u(p; t) \) is a solution to the initial value problem posed on the previous page and indeed it is the unique such solution.
When we allow the mutation rates to be positive, the initial value problem becomes

\[
\dot{u}(p; t) = (Np(1 - p) + N(1 - p)\mu_{21})\{u(p + 1/N; t) - u(p; t)\} + (Np(1 - p) + Np\mu_{12})\{u(p - 1/N; t) - u(p; t)\},
\]

again subject to the initial condition \(u(p; 0) = p\). To solve this, we will reinterpret the quantity \(u(p; t)\) in terms of the sampling distribution of the population. Suppose that we sample a single individual from the population at time \(t\). If \(G_t\) denotes the genotype of this individual, then

\[
P(G_t = A_1 | p_t = x) = x
\]

since the probability of sampling an allele is equal to its frequency. However, since we don’t know the value of \(p_t\) for \(t > 0\), we instead calculate

\[
P(G_t = A_1 | p_0 = p) = \sum_x P(G_t = A_1 | p_t = x)P(p_t = x | p_0 = p) = \sum_x xP(p_t = x | p_0 = p) = u(p; t).
\]
We then reason as follows.

- The probability that an individual randomly sampled from the population at time $t$ carries the genotype $A_1$ will depend on both the genotype of the unique ancestor of that individual that was alive at time 0 and on the number of mutations that occurred along the lineage leading from that ancestor to the sampled individual.

- The probability that the ancestor carried the $A_1$ genotype is just $p$.

- If we let $G_s$ denote the genotype of the ancestor of the sampled individual that was alive at time $s$, then $G = (G_s : 0 \leq s \leq t)$ is a two-state Markov chain with initial distribution $p_0 = (p, 1 - p)$ and transition rates $\mu_{12}$ for transitions from $A_1$ to $A_2$ and $\mu_{21}$ for transitions from $A_2$ to $A_1$.

Assuming that this reasoning is correct, we can use the results from our previous example to calculate that

$$u(p; t) = \mathbb{P}(G_t = A_1 | p_0 = p) = \frac{\mu_{21}}{\mu} + \frac{p \mu_{12} - (1 - p) \mu_{21}}{\mu} e^{-\mu t}$$

where we have introduced $\mu = \mu_{12} + \mu_{21}$. 
To verify that this is indeed the desired solution, all we need to do is to confirm that it solves the relevant initial value problem. To this end, we calculate that

\[
\dot{u}(p; t) = -(p\mu_{12} - (1 - p)\mu_{21})e^{-\mu t}
\]

\[
u(p + 1/N; t) - u(p; t) = \frac{1}{N}e^{-\mu t}
\]

\[
u(p - 1/N; t) - u(p; t) = -\frac{1}{N}e^{-\mu t}
\]

which gives

\[
\begin{align*}
(Np(1 - p) + N(1 - p)\mu_{21})\{u(p + 1/N; t) - u(p; t)\} \\
+ (Np(1 - p) + Np\mu_{12})\{u(p - 1/N; t) - u(p; t)\} \\
= (p(1 - p) + (1 - p)\mu_{21})e^{-\mu t} - (p(1 - p) + p\mu_{12})e^{-\mu t} \\
= -(p\mu_{12} - (1 - p)\mu_{21})e^{-\mu t}
\end{align*}
\]

as required.
Previously we saw that each of the jump times of a continuous-time Markov chain can be represented as the minimum of a collection of independent exponential random variables. Surprisingly, it turns out that these minima are themselves exponentially distributed.

**Lemma**

Let $\eta_j, j \in E$ be a countable collection of independent exponential random variables with rates $\lambda_j$ and assume that the sum of the rates is finite:

$$\lambda = \sum_{j \in E} \lambda_j < \infty.$$ 

If $\tau$ and $Y$ are defined by the identities

$$\tau = \inf_{j \in E} \eta_j \text{ and } Y = \arg \min_{j \in E} \eta_j,$$

then $\tau$ is an exponential random variable with parameter $\lambda$ and $Y$ is independent of $\tau$ with distribution

$$\mathbb{P}(Y = j) = \frac{\lambda_j}{\lambda}.$$
Proof: We can calculate the joint distribution of the variables $\tau$ and $Y$ as follows:

\[
\mathbb{P}(\tau > t, Y = j) = \mathbb{P}(t < \eta_j < \eta_k \quad \forall k \neq j)
\]
\[
= \int_t^\infty \lambda_j e^{-\lambda_j s} \mathbb{P}(\eta_k > s \quad \forall k \neq j) \, ds
\]
\[
= \lambda_j \int_t^\infty e^{-\lambda_j s} \prod_{k \neq j} e^{-\lambda_k s} \, ds
\]
\[
= \lambda_j \int_t^\infty e^{-\lambda s} \, ds
\]
\[
= \left( \frac{\lambda_j}{\lambda} \right) e^{-\lambda t}
\]
\[
= \mathbb{P}(Y = j) \cdot \mathbb{P}(\tau > t).
\]

This shows that $\tau$ and $Y$ are independent (since the joint distribution factors into a product of marginal distribution function) and that $\tau$ and $Y$ have the distributions claimed above. □
This lemma leads to an alternative method for simulating continuous-time Markov chains widely known as the **Gillespie algorithm**. Recall that we use the variables $J_1 < J_2 < \cdots$ to denote the **jump times** of the Markov chain.

- Sample a state $i_0 \in E$ with probability $\nu_{i_0}$ and set $X_0 = i_0$ and $J_0 = 0$; this is the initial state of the chain.
- Conditional on $X_{J_n} = i$ and $q_{ii} \neq 0$, generate a pair of independent random variables $\tau_{n+1}$ and $Y_{n+1}$, where $\tau_{n+1}$ is exponentially distributed with rate $|q_{ii}|$ and
  \[
  \mathbb{P}(Y_{n+1} = j) = \frac{q_{ij}}{|q_{ii}|}.
  \]
  If $q_{ii} = 0$, then $X$ is absorbed into the state $i$ and we set $X_{J_{n+s}} = i$ for all $s > 0$.
- Let $J_{n+1} = J_n + \tau_{n+1}$ be the time of the $n+1$’st jump and define $X$ on the interval $[J_n, J_{n+1}]$ by setting
  \[
  X_t = \begin{cases} 
  i & \text{if } t \in [J_n, J_{n+1}) \\
  Y_{n+1} & \text{if } t = J_{n+1}.
  \end{cases}
  \]
Remarks on the Gillespie algorithm:

- This algorithm was popularized by Dan Gillespie in the 1970's as a method for simulating stochastic chemical reactions. However, it was known to William Feller as early as 1940.
- The advantage of this approach is that rather than generating multiple exponential random variables, one for each possible transition out of the current state, we only need to generate two random variables, one for the holding time and one for the next state.
- The exponentially distributed holding time can be generated by first sampling a standard uniform random variable $U$ and then transforming it into

$$
\tau = -\frac{\ln(U)}{|q_{ii}|}.
$$

- The next state $Y_{n+1}$ can be sampled by dividing the unit interval $[0, 1]$ into disjoint subintervals of lengths $q_{i1}/|q_{ii}|$, $q_{i2}/|q_{ii}|$, … and then choosing one of these by generating a second standard uniform random variable $V$ that is independent of $U$. 
Another consequence of the lemma is that the sequence of states visited by a continuous-time Markov chain $X$ is itself a discrete-time Markov chain called the **jump chain** of $X$. To describe this process, let

$$q_i \equiv -q_{ii} = \sum_{j \neq i} q_{ij}, \quad i \in E$$

be the total rate at which the process $X$ leaves state $i$. For the moment, let us assume that $q_i > 0$ for every $i \in E$. To define the jump chain, let $J_n$ be the time of the $n$th jump by $X$ and define the discrete-time process $Y = (Y_n : n \geq 0)$ by setting $Y_0 = X_0$ and

$$Y_n = X_{J_n},$$

for all $n \geq 1$. Then $Y$ is a discrete-time Markov chain with the same initial distribution as $X$ and transition matrix $P = (p_{ij})$, where $p_{ii} = 0$ for all $i \in E$ and

$$p_{ij} = \frac{q_{ij}}{q_i},$$

for all $j \neq i$. Notice that $Y$ is not a typical discrete-time Markov chain insofar as it always moves to a new state at every time step.
If \( q_{ii} = 0 \) for some \( i \in E \), then the continuous-time chain \( X \) has absorbing states and the sequence of states visited by \( X \) may be finite. Let \( A \) be the set of all such absorbing states. In this case, the jump chain \( Y \) is said to be a **stopped Markov chain** and is defined only up to a random time

\[
T = \inf\{n \geq 0 : X_{J_n} \in A\},
\]

when \( X \) is absorbed by some state in \( A \). If \( X \) is not absorbed, then we set \( T = \infty \). In either case, we can write \( Y = (Y_n : 0 \leq n \leq T) \) where \( Y_T \in A \) if \( T < \infty \). The transition matrix \( P \) of the stopped chain is slightly different from that on the preceding slide:

\[
p_{ij} = \begin{cases} 
1 & \text{if } j = i \in A \\
0 & \text{if } j \neq i \in A \\
0 & \text{if } j = i \notin A \\
q_{ij}/q_i & \text{if } j \neq i \notin A 
\end{cases}
\]

Notice that with this convention, the jump chain \( Y \) has the same set of absorbing states as \( X \).
Because $X$ and $Y$ visit the same sequence of states, we can use our previous results on discrete-time Markov processes to calculate the absorption probabilities for $X$. Suppose that $a \in A$ is an absorbing state for $X$ and $Y$ and let $h = (h_1, h_2, \cdots)$ be the absorption probabilities for these two processes:

$$h_i = \mathbb{P}(X_t = a \text{ for some } t \geq 0|X_0 = i) = \mathbb{P}(Y_n = a \text{ for some } n \geq 0|Y_0 = i).$$

These probabilities are identical because $X$ absorbs in state $a$ if and only if $Y$ absorbs in state $a$. Previously we showed that the vector $h$ is the minimal non-negative solution to the following linear system of equations:

$$\begin{cases} h_i = 1 & \text{if } i = a \\ h_i = \sum_{j \in E} p_{ij} h_j & \text{if } i \neq a. \end{cases}$$

Furthermore, we showed that $h$ is a right-eigenvector of $P$ corresponding to the eigenvalue 1:

$$Ph = h.$$
We can also express the absorption probabilities in terms of the rate matrix $Q$. Indeed, if we multiply the identity $h_i = \sum_j p_{ij} h_j$ by $q_i$ and then subtract $q_i h_i$ from both sides, we arrive at the following system of linear equations:

$$
\begin{align*}
    h_a &= 1 \\
    \sum_{j \in E} q_{ij} h_j &= 0.
\end{align*}
$$

Since the second equation holds even when $i = a$ (since $q_{aj} = 0$ for all $j \in E$ if $a$ is an absorbing state), it follows that $h$ is a right-eigenvector of the rate matrix with eigenvalue 0:

$$Qh = 0$$

where $0$ is a column vector indexed by $E$ with all entries equal to 0.
First Passage Times

Suppose that $X$ is a continuous-time Markov chain with rate matrix $Q$. We will define the **first passage time** (also called the **hitting time**) of a state $k \in E$ to be the random variable

$$T_k \equiv \inf\{t \geq 0 : X_t = k\},$$

with $T_k = \infty$ if $X_t \neq k$ for all $t \geq 0$. As long as the process $X$ has right-continuous sample paths, i.e.,

$$\mathbb{P}\left(X_t = \lim_{s \to 0^+} X_{t+s} \forall t \geq 0\right) = 1,$$

we can interpret $T_k$ as the first time that the process $X$ occupies the state $k$. This will be true, for example, if $Q$ is bounded and we construct $X$ using the Gillespie algorithm described previously.
If we write \(( T_k | X_0 = i )\) for the hitting time of \( k \) conditional on \( X_0 = i \), then we can use the fact that \( X \) is a Markov process to show that the following identity holds whenever \( i \) is not an absorbing state (so that \( q_i > 0 \)):

\[
( T_k | X_0 = i ) \overset{d}{=} \eta_i + \sum_{j \neq i} \frac{q_{ij}}{q_i} ( T_k | X_0 = j ),
\]

where \( \eta_i \) is an exponentially distributed random variable with rate \( q_i \) and the variables \(( T_k | X_0 = j ), j \in E \) are independent of each other and of \( \eta_i \). In other words, the time required to hit \( k \) starting in state \( i \) is equal to the initial holding time in state \( i \) plus the time required to hit state \( k \) starting in the next state visited by \( X \).

**Notation:** The expression \( X \overset{d}{=} Y \) means that \( X \) and \( Y \) have the same distribution, not necessarily that \( X = Y \). For example, if \( U \) and \( V \) are independent standard uniform random variables, then \( U \overset{d}{=} V \), but \( \mathbb{P}(U = V) = 0 \). On the other hand, identity in distribution does imply that \( \mathbb{E}[f(X)] = \mathbb{E}[f(Y)] \) for all bounded functions \( f \).
If we fix \( k \in E \) and let \( t_i = \mathbb{E}[T_k | X_0 = i] \) denote the mean time for the first passage from \( i \) to \( k \), then by taking expectations of both sides of the identity on the preceding slide we see that the vector \( t = (t_1, t_2, \cdots) \) satisfies the following system of linear equations

\[
  t_i = \frac{1}{q_i} + \sum_{j \neq i} \frac{q_{ij}}{q_i} t_j, \quad i \neq k
\]

with \( t_k = 0 \). Furthermore, if we multiply both sides of this equation by \( q_i \) and then subtract \( q_i t_i = -q_{ii} t_i \) and 1 from both sides, we can rewrite this system in the following equivalent form:

\[
  \sum_{j \in E} q_{ij}(t_j - t_i) = -1, \quad i \neq k.
\]
**Example: Simple Random Walk with Reflecting Boundaries** Suppose that $X$ is the continuous-time Markov chain with state space $E = \{0, 1, \cdots, L\}$ and transition rates

$$q_{ij} = \begin{cases} 
\lambda & \text{if } |j - i| = 1, i, j \in E \\
-\lambda & \text{if } j = i \in \{0, L\} \\
-2\lambda & \text{if } j = i \in \{1, \cdots, L - 1\} \\
0 & \text{otherwise}.
\end{cases}$$

In other words, the process executes a symmetric random walk with reflecting boundaries at 0 and $L$ and exponentially distributed jump times. Let $T_0$ denote the first passage time to 0 and let $t_i = \mathbb{E}[T_0|X_0 = i]$ be the expected value of this time when the process starts in state $i$. Then $t_0 = 0$ and the values $t_1, \cdots, t_L$ satisfy the following system of equations

$$\lambda(t_{i+1} - t_i) + \lambda(t_{i-1} - t_i) = -1, \quad 1 \leq i \leq L - 1,$$

$$\lambda(t_{L-1} - t_L) = -1.$$
To solve for $t_i$, we first rewrite this system of equations in terms of the auxiliary variables $\Delta_i = t_{i+1} - t_i$, $i = 0, \cdots, L - 1$:

$$
\lambda(\Delta_i - \Delta_{i-1}) = -1, \quad 1 \leq i \leq L - 1,
$$

$$
-\lambda\Delta_{L-1} = -1.
$$

This can be solved recursively, giving,

$$
\Delta_{L-1} = \frac{1}{\lambda}
$$

$$
\Delta_{L-2} = \Delta_{L-1} + \frac{1}{\lambda} = \frac{2}{\lambda}
$$

$$
\Delta_{L-3} = \Delta_{L-2} + \frac{1}{\lambda} = \frac{3}{\lambda}
$$

$$
\vdots
$$

$$
\Delta_i = \frac{L - i}{\lambda}.
$$
Next, we use the fact that $t_0 = 0$ and $t_{i+1} = t_i + \Delta_i$ to recursively solve for $t_i$:

\[
\begin{align*}
    t_0 &= 0 \\
    t_1 &= \Delta_0 = \frac{L}{\lambda} \\
    t_2 &= t_1 + \Delta_1 = \frac{L}{\lambda} + \frac{L - 1}{\lambda} \\
    \ldots & \ldots \\
    t_i &= t_{i-1} + \Delta_{i-1} = \sum_{j=0}^{i-1} \frac{L-j}{\lambda} \\
    &= \frac{1}{\lambda} \left\{ iL - \binom{i}{2} \right\} \\
    &= \frac{i}{\lambda} \left( L - \frac{i-1}{2} \right).
\end{align*}
\]
By symmetry, it follows that for any \( i, j \in E \) the expected first passage time \( t_{i,j} \) from state \( i \) to state \( j \) is equal to the expected first passage time from state \( |i - j| \) to 0 and so

\[
t_{i,j} = \frac{|i - j|}{\lambda} \left( L - \frac{|i - j| - 1}{2} \right),
\]

which grows linearly in the size \( L \) of the region and nearly linearly in the distance between the initial and final states. Furthermore, by taking the limit as \( L \to \infty \), we see that the first passage times for one-dimensional random walks on \( \mathbb{Z} \) have infinite expectations.
The recurrence equation for the first passage times can also be used to derive a system of linear equations satisfied by the moment generating functions of these variables. For each \( i \in E \), define the function \( \phi_i : [0, \infty) \to [0, 1] \) by

\[
\phi_i(\omega) = \mathbb{E} \left[ e^{-\omega T_{k\mid X_0=i}} \right] = \sum_{n=0}^{\infty} \frac{1}{n!} \mathbb{E} [T_k^n] (-\omega)^n.
\]

\( \phi_i \) is said to be the moment generating function of \( (T_k \mid X_0 = i) \) and determines the distribution of this random variable uniquely; indeed, \( \phi_i \) is also the Laplace transform of this distribution. Using the fact that the variables on the right-hand side of the recurrence equation are independent, we have

\[
\phi_i(\omega) = \mathbb{E} \left[ e^{-\omega \eta_i} \right] \sum_{j \neq i} \frac{q_{ij}}{q_i} \mathbb{E} [e^{-\omega (T_k \mid X_0=j)}] \\
= \frac{q_i}{q_i + \omega} \sum_{j \neq i} \frac{q_{ij}}{q_i} \phi_j(\omega) \\
= \frac{1}{q_i + \omega} \sum_{j \neq i} q_{ij} \phi_j(\omega),
\]

with \( \phi_k(\omega) = 1 \) for all \( \omega \). Depending on \( Q \), it may be possible to solve for the \( \phi_i \) and then invert the Laplace transform to recover the distribution of \( (T_k \mid X_0 = i) \).
Stationary Distributions

We begin with the main definition.

**Definition**

Let $X$ be a continuous-time Markov chain with rate matrix $Q$ and state space $E$. A distribution $\pi = (\pi_1, \pi_2, \cdots)$ on $E$ is said to be a *stationary distribution* for $X$ if

$$\pi Q = 0$$

i.e., $\pi$ is a left eigenvector for $Q$ corresponding to 0.

Observe that if $X_0 \sim \pi$, where $\pi$ is a stationary distribution, then the marginal distribution $p_t$ of the chain at time $t$ is

$$p_t = \pi e^{Qt} = \pi \sum_{n=0}^{\infty} \frac{1}{n!} Q^n t^n = \pi \left( I + Qt + \frac{1}{2} Q^2 t^2 + \cdots \right) = \pi,$$

for every $t \geq 0$. 
The condition $\pi Q = 0$ can also be understood in terms of conservation of mass. If we write this condition in terms of the individual components of $\pi$ and $Q$, we obtain

$$0 = \sum_{j \in E} \pi_j q_{ji}$$

$$= \sum_{j \neq i} \pi_j q_{ji} + \pi_i q_{ii}$$

$$= \sum_{j \neq i} \pi_j q_{ji} - \pi_i q_i$$

which asserts that, at equilibrium, the total rate at which probability mass flows into state $i$ must be balanced by the rate at which probability mass flows out of that state:

$$\sum_{j \neq i} \pi_j q_{ji} = \pi_i q_i.$$ 

In particular, a stationary distribution $\pi$ is also a stationary solution of the Kolmogorov forward equation.
Suppose that $X$ is a continuous-time Markov chain with rate matrix $Q = (q_{ij})$ and let $Y$ be its jump chain with transition matrix $P = (p_{ij})$. Here we will assume that $X$ has no absorbing states so that $p_{ij} = q_{ij}/q_i$ if $j \neq i$ and $p_{ii} = 0$.

As a general rule, $X$ and $Y$ will not have the same stationary distributions, although their stationary distributions will be related to one another. For example, if $\pi$ is a stationary distribution for $X$, then we can define another probability distribution $\mu = (\mu_1, \mu_2, \cdots)$ on $E$ by setting

$$\mu_i = \frac{q_i \pi_i}{\bar{q}}$$

provided that the average jump rate

$$\bar{q} = \sum_{j \in E} \pi_j q_j$$

is finite. Division by $\bar{q}$ guarantees that $\mu$ is a probability distribution on $E$. 
I claim that $\mu$ is a stationary distribution for the jump chain $Y$. To verify this, we need to show that $\mu P = \mu$, which we can do componentwise:

$$(\mu P)_i = \sum_{j \in E} \mu_j p_{ji}$$

$$= \sum_{j \neq i} \left( \frac{q_j \pi_j}{\bar{q}} \right) \left( \frac{q_{ji}}{q_j} \right)$$

$$= \frac{1}{\bar{q}} \sum_{j \neq i} \pi_j q_{ji}$$

$$= \frac{1}{\bar{q}} \pi_i q_i = \mu_i.$$

Notice that the two distributions will be equal only if $q_i = \bar{q}$ for all $i \in E$, i.e., only if $X$ jumps out of each state at the same rate. Otherwise, the probabilities $\pi_i$ will be inversely proportional to the jump rates $q_i$, with $X$ spending proportionately less time in states with high jump rates than the jump process $Y$. 

**Example:** We previously showed that if $X$ is the two-state continuous-time Markov chain with rate matrix

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

and initial distribution $\nu = (\nu_1, \nu_2)$, then the marginal distribution $p(t)$ of $X$ at time $t \geq 0$ is equal to

$$p_1(t) = \frac{b}{a + b} + \frac{\nu_1 a - \nu_2 b}{a + b} e^{-(a+b)t},$$

$$p_2(t) = \frac{a}{a + b} - \frac{\nu_1 a - \nu_2 b}{a + b} e^{-(a+b)t}.$$

Letting $t \to \infty$, we see that if $a, b > 0$, then $p(t)$ converges to the distribution

$$\pi = \begin{pmatrix} b \\ a \end{pmatrix} \frac{1}{a + b},$$

$$\begin{pmatrix} a \\ a + b \end{pmatrix},$$

which is the unique stationary distribution for this Markov chain.
Example: Recall that the rate matrix and the corresponding transition matrices for the Jukes-Cantor model are

\[
Q = \begin{pmatrix}
-3\mu & \mu & \mu & \mu \\
\mu & -3\mu & \mu & \mu \\
\mu & \mu & -3\mu & \mu \\
\mu & \mu & \mu & -3\mu \\
\end{pmatrix},
\]

and

\[
P(t) = e^{Qt} = \frac{1}{4} \begin{pmatrix}
1 + 3e^{-4\mu t} & 1 - e^{-4\mu t} & 1 - e^{-4\mu t} & 1 - e^{-4\mu t} \\
1 - e^{-4\mu t} & 1 + 3e^{-4\mu t} & 1 - e^{-4\mu t} & 1 - e^{-4\mu t} \\
1 - e^{-4\mu t} & 1 - e^{-4\mu t} & 1 + 3e^{-4\mu t} & 1 - e^{-4\mu t} \\
1 - e^{-4\mu t} & 1 - e^{-4\mu t} & 1 - e^{-4\mu t} & 1 + 3e^{-4\mu t} \\
\end{pmatrix}.
\]

Letting \( t \to \infty \), we see that \( p_{xy}(t) \to 1/4 \) for every \( x, y \in \{T, C, A, G\} \) and that \( \pi = (1/4, 1/4, 1/4, 1/4) \) is a stationary distribution for \( Q \).
In the two examples just presented, we saw that the transition probabilities tended to the unique stationary distribution of the Markov chain at large times and irrespective of the initial state: \( p_{ij}(t) \rightarrow \pi_j \) as \( t \rightarrow \infty \). Although not all continuous-time Markov chains have this property, the next theorem asserts that a fairly large class of chains do behave in this manner.

**Theorem**

Let \( X \) be a continuous-time Markov chain with bounded rate matrix \( Q \) and suppose that the jump chain \( Y \) is irreducible and that \( X \) has a stationary distribution \( \pi \). Then for all states \( i, j \in E \),

\[
\lim_{t \to \infty} p_{ij}(t) = \pi_j.
\]

In particular, it follows that \( \pi \) is the unique stationary distribution of \( X \) and that the marginal distributions of the variables \( X_t \) tend to \( \pi \) as \( t \rightarrow \infty \) for all initial distributions of \( X_0 \). Such processes are said to be **ergodic** and are further described by the next theorem.
Theorem

(Ergodic Theorem for Markov Chains.) Suppose that $X$ is a continuous-time Markov chain with initial distribution $\nu$ which satisfies the conditions of the previous theorem. Then, for each $i \in E$

$$\lim_{t \to \infty} \frac{1}{t} \int_0^t 1\{X_s = i\} \, ds = \pi_i \quad \text{a.s.,}$$

i.e., the limit exists and is equal to $\pi_i$ with probability 1. Furthermore, if $f : E \to \mathbb{R}$ is a bounded function, then

$$\lim_{t \to \infty} \frac{1}{t} \int_0^t f(X_s = i) \, ds = \bar{f} \quad \text{a.s.,}$$

where

$$\bar{f} = \sum_{i \in E} \pi_i f(i)$$

is the expected value of $f$ averaged over the stationary distribution $\pi$.

Remark: In effect, this theorem asserts that the time averages of a function $f$ along a sample path of the chain tend to the spatial average of the function over the state space $E$ with respect to the stationary distribution $\pi$. 
Continuous-Time Birth-Death Processes

A continuous-time Markov chain $X$ is said to be a continuous-time birth-death process if the state space is $E = \{0, 1, \cdots, N\}$, where $N$ may be finite or infinite, and the rate matrix $Q$ is tridiagonal, i.e.,

$$q_{ij} = \begin{cases} 
\lambda_i & \text{if } j = i + 1 \\
\mu_i & \text{if } j = i - 1 \\
-(\mu_i + \lambda_i) & \text{if } j = i \\
0 & \text{if } |j - i| > 1, 
\end{cases}$$

with $\mu_0 = \lambda_N = 0$. Often we think of $\lambda_i$ and $\mu_i$ as the total birth rate and total death rate, respectively, in a population containing $i$ individuals. As with their discrete-time cousins, the simple structure of the rate matrix of a continuous-time birth-death processes can often be exploited when performing calculations.
By way of example, suppose that $X$ is a continuous-time birth-death process and that

$$
\mu_i > 0 \text{ and } \lambda_{i-1} > 0 \text{ for } i = 1, \cdots, N,
$$

and

$$
\sum_{i=1}^{N} \frac{\lambda_0 \lambda_1 \cdots \lambda_{i-1}}{\mu_1 \mu_2 \cdots \mu_i} < \infty.
$$

If these conditions are satisfied, then it can be shown that $X$ has a unique stationary distribution $\pi = (\pi_0, \pi_1, \pi_2, \cdots)$ on $E$. Since $\pi Q = 0$, it follows that these probabilities must satisfy the following system of equations:

\begin{align*}
0 &= -\lambda_0 \pi_0 + \mu_1 \pi_1 \\
0 &= \lambda_{i-1} \pi_{i-1} - (\lambda_i + \mu_i) \pi_i + \mu_{i+1} \pi_{i+1}, \quad 1 \leq i < N \\
0 &= \lambda_{N-1} \pi_{N-1} - \mu_N \pi_N \quad \text{(if } N < \infty\text{).}
\end{align*}
These equations can be solved recursively as follows. Starting with $i = 0$, we have

$$\pi_1 = \frac{\lambda_0}{\mu_1} \pi_0$$

Then with $i = 1$,

$$\pi_2 = \frac{1}{\mu_2} \left\{ (\lambda_1 + \mu_1) \pi_1 - \lambda_0 \pi_0 \right\}$$
$$= \frac{1}{\mu_2} \lambda_1 \pi_1$$
$$= \frac{\lambda_0 \lambda_1}{\mu_1 \mu_2} \pi_0.$$

Continuing in this way, one finds that for general $i = 1, \cdots, N$,

$$\pi_i = \frac{\lambda_0 \lambda_1 \cdots \lambda_{i-1}}{\mu_1 \mu_2 \cdots \mu_i} \pi_0.$$
For $\pi$ to be a probability distribution on $E$, these quantities must sum to 1:

$$1 = \sum_{i=0}^{N} \pi_i = \pi_0 \left( 1 + \sum_{i=1}^{N} \frac{\lambda_0 \lambda_1 \cdots \lambda_{i-1}}{\mu_1 \mu_2 \cdots \mu_i} \right)$$

which gives

$$\pi_0 = \left( 1 + \sum_{i=1}^{N} \frac{\lambda_0 \lambda_1 \cdots \lambda_{i-1}}{\mu_1 \mu_2 \cdots \mu_i} \right)^{-1} > 0$$

since our second condition on the birth and death rates guarantees that the sum inside the parentheses is finite even if $N$ is infinite. This will be true, for example, if there is an $\epsilon > 0$ and a positive integer $K$ such that $\lambda_i < (1 - \epsilon)\mu_i$ for all $i \geq K$, i.e., at sufficiently high densities, the birth rates need to be geometrically less than the death rates. Otherwise, the population might grow without bound, in which case there need not be a stationary distribution.
As with discrete-time Markov chains, if we run a continuous-time Markov chain \( X = (X_t : t \in \mathbb{R}) \) backwards in time, then the time-reversed process \( \hat{X} = (\hat{X}_t : t \in \mathbb{R}) \) obtained by setting

\[
\hat{X}_t = X_{T-t}
\]

is also a continuous-time Markov process, although it need not be time-homogeneous. This is so because the requirement that the past and future be conditionally independent given the present remains true when the past and future are interchanged. Furthermore, if for every \( t \), \( X_t \sim \pi \) where \( \pi \) is a stationary distribution, then it can be shown that \( \hat{X} \) is also a stationary Markov chain with rate matrix \( \hat{Q} = (\hat{q}_{ij}) \), where

\[
\hat{q}_{ij} = \begin{cases} 
q_{ji} \left( \frac{\pi_j}{\pi_i} \right) & \text{if } \pi_i > 0 \\
0 & \text{if } \pi_i = 0.
\end{cases}
\]
If all of the probabilities of the stationary distribution are positive and \( \Pi = \text{diag}(\pi) \) is a diagonal matrix with diagonal entries \( \Pi_{ii} = \pi_i \), then the relationship between \( \hat{Q} \) and \( Q \) can also be expressed as

\[
\hat{Q} = \Pi^{-1} Q^T \Pi,
\]

where \( Q^T = (q_{ji}) \) is the transpose of \( Q \). In general, \( \hat{Q} \neq Q \) and so the transition rates of the time-reversed process will usually differ from those of the forward process. With discrete-time chains we saw that detailed balance is a sufficient condition for the transition matrix to be invariant under time reversal and a similar result holds in the continuous-time case prompting the following definition.

**Definition**

A continuous-time Markov chain \( X \) with rate matrix \( Q \) is said to satisfy **detailed balance** with respect to a stationary distribution \( \pi \) if for all \( i, j \in E \),

\[
\pi_i q_{ij} = \pi_j q_{ji}.
\]
If detailed balance holds with respect to a stationary distribution, then a simple calculation shows that the transition rates of the time-reversed process are

$$\hat{q}_{ij} = \frac{\pi_j q_{ji}}{\pi_i} = \frac{\pi_i q_{ij}}{\pi_i} = q_{ij}$$

and so $\hat{Q} = Q$, meaning that the statistics of the time-reversed process are indistinguishable from those of the forwards-in-time process. In such cases, the process $X$ is said to be **reversible**.
Example: The General Time-Reversible Model of DNA Substitution

Previously we introduced a very simple model for the evolution of DNA sequences known as the Jukes-Cantor model which made the following assumptions:

- All substitutions occur at equal rates, e.g., $A \rightarrow T$ occurs at the same rate as $C \rightarrow A$.
- The equilibrium frequencies of the four nucleotides are $(1/4, 1/4, 1/4, 1/4)$.

Although this model was useful in the 1970’s when both the sequence data and the computational resources available for analysis of that data were quite limited, subsequent work has revealed that both of the assumptions underpinning JC69 are almost always badly violated.
Because statistical phylogenetics is usually done using unrooted trees, there is an advantage to working with substitution models that are reversible continuous-time Markov chains on the set \{T, C, A, G\}. The most general such model (known as the **General Time-Reversible Model** or GTR for short) has the following rate matrix

\[
Q = \begin{pmatrix}
\circ & \pi_C \alpha & \pi_A \beta & \pi_G \gamma \\
\pi_T \alpha & \circ & \pi_A \delta & \pi_G \epsilon \\
\pi_T \beta & \pi_C \delta & \circ & \pi_G \eta \\
\pi_T \gamma & \pi_C \epsilon & \pi_A \eta & \circ
\end{pmatrix},
\]

where

- \(\pi = (\pi_T, \pi_C, \pi_A, \pi_G)\) is the stationary distribution for \(Q\);
- \(\alpha, \beta, \gamma, \delta, \epsilon, \eta\) are six free parameters that can be estimated using sequence data;
- the circles on the diagonal indicate that these entries should be chosen so that the row sums are zero.

Since this matrix has been formulated so that \(q_{yx} = q_{xy} \pi_x / \pi_y\), it is clear that detailed balance is satisfied by the distribution \(\pi\).