MTH 428/528

Lectures 12-15

Yevgeniy Kovchegov
Oregon State University
Topics:

- Fixation times for Moran model
- Martingales
**Wright-Fisher Model.**

**Model:** Let $X_t$ denote the number of alleles $A$ in $t$-th generation. Thus, there are $2N - X_t$ of alleles $a$.

For each of the $2N$ copies of the locus in $(t + 1)$-st generation, one of the $2N$ alleles in $t$-th generation is selected uniformly at random. Thus, in these $2N$ Bernoulli trials, allele $A$ comes up with probability

$$p = \frac{X_t}{2N}$$

and allele $a$ comes up with probability $1 - p$.

Hence, given that we know $X_t$, variable $X_{t+1}$ representing the number of alleles $A$ in $(t+1)$-st generation is a **Binomial random variable** with $2N$ trials and $p = \frac{X_t}{2N}$, i.e.

$$P(X_{t+1} = k \mid X_t = m) = \binom{2N}{k} p^k (1 - p)^{2N-k}, \quad \text{where } p = \frac{m}{2N}. $$
Wright-Fisher Model.

Hence, given that we know $X_t$, variable $X_{t+1}$ representing the number of alleles $A$ in $(t+1)$-st generation is a **Binomial random variable** with $2N$ trials and $p = \frac{X_t}{2N}$, i.e.

$$P(X_{t+1} = k \mid X_t = m) = \binom{2N}{k} p^k (1 - p)^{2N-k},$$

where $p = \frac{m}{2N}$.

**Urn model and fixation time:**
Moran Model.

Consider $N$ individuals, $n = 2N$ copies of the locus.

- Wright-Fisher Model: non-overlapping generations (e.g. annual plants).

- Moran Model (1958): one individual changes at a time (we have $n = 2N$ ‘individuals’).

There, $X_t$ denote the number of alleles $A$ at time $t$.

\[
P(X_{t+1} = i + 1 \mid X_t = i) = \frac{i}{n} \cdot \frac{n-i}{n}
\]

\[
P(X_{t+1} = i - 1 \mid X_t = i) = \frac{n-i}{n} \cdot \frac{i}{n}
\]

State space: $S = \{0, 1, 2, \ldots, n\}$. 
Fixation times for Moran model.

Consider a birth-and-death chain $X_t$ on $S = \{0, 1, \ldots, n\}$ with forward probabilities and backward probabilities denoted respectively $p_j$ and $q_j$. Our goal is to find fixation time $T_f$ which is defined as the following first hitting time:

$$T_f = \min\{t \geq 0 : X_t = 0 \text{ or } n\}.$$

We let

$$\varphi(j) = E[T_f \mid X_0 = j],$$

and write the following recurrence equation:

$$\begin{cases} 
\varphi(j) = 1 + q_j \varphi(j - 1) + r_j \varphi(j) + p_j \varphi(j + 1) & \text{for } j = 1, 2, \ldots, n - 1 \\
\varphi(0) = \varphi(n) = 0. 
\end{cases}$$
Fixation times for Moran model: discussion.

We let \( \ell(1) = 0 \) and for \( j \geq 2 \), \( \ell(j) := \sum_{k=1}^{j-1} \frac{1}{k} \), and obtained

\[
\varphi(j) = n^2 \left[ \frac{j}{n} (\ell(n) - \ell(j)) + \left(1 - \frac{j}{n}\right) (\ell(n) - \ell(n - j + 1)) \right].
\]

Observe that

\[
\ell(j) := \sum_{k:1 \leq k \leq j-1} \frac{1}{k} = \int_{1}^{j} \frac{1}{x} \, dx + C_j = \ln(j) + C_j,
\]

where \( 0 < C_j < \gamma = 0.57721 \ldots \) (the Euler constant).
Fixation times for Moran model: discussion.

\[ \ell(j) := \sum_{k:1 \leq k \leq j-1} \frac{1}{k} = \int_{1}^{j} \frac{1}{x} dx + C_j = \ln(j) + C_j, \]

where \(0 < C_j < \gamma = 0.57721\ldots\) (the Euler constant).

Source: Wikipedia.org
Fixation times for Moran model: discussion.

We let $\ell(1) = 0$ and for $j \geq 2$, $\ell(j) := \sum_{k=1}^{j-1} \frac{1}{k}$, and obtained

$$\varphi(j) = n^2 \left[ \frac{j}{n} (\ell(n) - \ell(j)) + \left(1 - \frac{j}{n}\right) (\ell(n) - \ell(n - j + 1)) \right].$$

Observe that

$$\ell(j) := \sum_{k:1 \leq k \leq j-1} \frac{1}{k} = \int_{1}^{j} \frac{1}{x} dx + C_j = \ln(j) + C_j,$$

where $0 < C_j < \gamma = 0.57721\ldots$ (the Euler constant). Hence,

$$\varphi(j) = -n^2 \left[ \frac{j}{n} \ln \frac{j}{n} + \left(1 - \frac{j}{n}\right) \ln \left(1 - \frac{j-1}{n}\right) + \text{Error} \right].$$
Fixation times for Moran model: discussion.

Observe that

\[ \ell(j) := \sum_{k:1 \leq k \leq j-1} \frac{1}{k} = \int_1^j \frac{1}{x} dx + C_j = \ln(j) + C_j, \]

where \(0 < C_j < \gamma = 0.57721\ldots\) (the Euler constant). Hence,

\[ \varphi(j) = -n^2 \left[ \frac{j}{n} \ln \frac{j}{n} + \left(1 - \frac{j}{n}\right) \ln \left(1 - \frac{j-1}{n}\right) + \text{Error} \right]. \]

If we let \(x = \frac{j}{n}\) be the fraction of \(A\) alleles, then for \(x \in [\epsilon, 1-\epsilon]\), the constant term goes to zero as \(n \to \infty\), and

\[ E[T_f \mid X_0/n = x] = -n^2 \left[ x \ln(x) + (1-x) \ln(1-x) + o(1) \right]. \]
Fixation times for Moran model.

**Alternative approach:** Let \( x = \frac{i}{n} \) be the fraction of \( A \) alleles, and write
\[
\varphi(j) = n^2 \psi(x).
\]
Recall approximation by 2nd Taylor polynomial:
\[
\psi(x + h) = \psi(x) + \psi'(x)h + \frac{1}{2} \psi''(x)h^2 + O(h^3).
\]
Hence, for \( h = \frac{1}{n} \),
\[
\Delta \varphi(j) = n^2 \left[ \psi \left( x + \frac{1}{n} \right) - \psi(x) \right] = \psi'(x)n + \frac{1}{2} \psi''(x) + O \left( \frac{1}{n} \right),
\]
and for \( h = -\frac{1}{n} \),
\[
-\Delta \varphi(j-1) = n^2 \left[ \psi \left( x - \frac{1}{n} \right) - \psi(x) \right] = -\psi'(x)n + \frac{1}{2} \psi''(x) + O \left( \frac{1}{n} \right).
\]
Fixation times for Moran model.

**Alternative approach:** Let $x = \frac{i}{n}$ and write

$$\varphi(j) = n^2 \psi(x).$$

Recall approximation by 2nd Taylor polynomial: for $h = \frac{1}{n}$,

$$\Delta \varphi(j) = n^2 \left[ \psi \left( x + \frac{1}{n} \right) - \psi(x) \right] = \psi'(x)n + \frac{1}{2} \psi''(x) + O \left( \frac{1}{n} \right),$$

and for $h = -\frac{1}{n}$,

$$-\Delta \varphi(j-1) = n^2 \left[ \psi \left( x - \frac{1}{n} \right) - \psi(x) \right] = -\psi'(x)n + \frac{1}{2} \psi''(x) + O \left( \frac{1}{n} \right).$$

The recursion $\Delta \varphi(j) - \Delta \varphi(j - 1) = -\frac{1}{p_j}$ rewrites as

$$\psi''(x) = -\frac{1}{x(1-x)} + O \left( \frac{1}{n} \right) \quad \text{with} \quad \psi(0) = \psi(1) = 0.$$
Fixation times for Moran model.

**Alternative approach:** Let $x = \frac{j}{n}$ and write

$$\varphi(j) = n^2 \psi(x).$$

The recursion $\Delta \varphi(j) - \Delta \varphi(j - 1) = -\frac{1}{p_j}$ rewrites as

$$\psi''(x) = -\frac{1}{x(1-x)} + O \left( \frac{1}{n} \right)$$

with $\psi(0) = \psi(1) = 0$.

Solving a “close” equation

$$\tilde{\psi}''(x) = -\frac{1}{x(1-x)}$$

with $\tilde{\psi}(0) = \tilde{\psi}(1) = 0$

we obtain

$$\tilde{\psi}(x) = -\left[ x \ln x + (1 - x) \ln(1 - x) \right].$$
Fixation times for Moran model.

While $O\left(\frac{1}{n}\right)$ in the original differential equation would result in the additional error term,

$$
\psi(x) = -\left[ x \ln x + (1 - x) \ln(1 - x) + O\left(\frac{1}{n}\right) \right],
$$

once again arriving with

$$
E[T_f \mid X_0/n = x] = -n^2 \left[ x \ln(x) + (1 - x) \ln(1 - x) + O\left(\frac{1}{n}\right) \right].
$$

Indeed, let $d(x) = \psi(x) - \tilde{\psi}(x)$, then $d(0) = d(1) = 0$ and $d''(x) = O\left(\frac{1}{n}\right)$. Thus $d'(x) = d'(0) + O\left(\frac{1}{n}\right)$ and $d(x) = d(0) + d'(0)x + O\left(\frac{1}{n}\right)$. Next, letting $x = 1$, obtain $d'(0) = d(1) - d(0) + O\left(\frac{1}{n}\right) = O\left(\frac{1}{n}\right)$, and therefore

$$
\psi(x) = \tilde{\psi}(x) + d(x) = \tilde{\psi}(x) + d(0) + d'(0)x + O\left(\frac{1}{n}\right) = \tilde{\psi}(x) + O\left(\frac{1}{n}\right).
$$
**Martingales.**

**Definition.** A homogeneous Markov chain \( \{X_t\} \) is a martingale if

- \( E[|X_t|] < \infty \) for all \( t \geq 0 \), and

\[
E[X_{t+1} \mid X_t = x] = x.
\]

**Definition.** For a homogeneous Markov chain \( \{X_t\} \), a random variable \( \tau \) is a stopping time if for any \( t \geq 0 \), knowing \( X_0, X_1, \ldots, X_t \) (i.e. the trajectory of the process up to time \( t \)) is sufficient for determining whether the event \( \{\tau \leq t\} \) occurred or not.

**Optional Stopping Theorem.** Suppose a homogeneous Markov chain \( \{X_t\} \) is a martingale, and \( T \) is a stopping time with respect to \( X_t \). If \( P(T < \infty) = 1 \) and there is \( K > 0 \) such that \( |X_t| \leq K \) when \( t < T \), then

\[
E[X_T \mid X_0] = X_0.
\]
Martingales.

If Markov chain \( \{X_t\} \) is a martingale, then

\[
E[X_1 \mid X_0 = x] = x \quad \text{and} \quad E[X_2 \mid X_1 = y] = y.
\]

Then,

\[
E[X_2 \mid X_0 = x] = \sum_{y \in S} E[X_2 \mid X_1 = y, X_0 = x] P(X_1 = y \mid X_0 = x)
\]

\[
= \sum_{y \in S} E[X_2 \mid X_1 = y] P(X_1 = y \mid X_0 = x)
\]

\[
= \sum_{y \in S} y P(X_1 = y \mid X_0 = x) = E[X_1 \mid X_0 = x] = x
\]

So, \( E[X_2 \mid X_0 = x] = x \), and iterating the argument, we obtain

\[
E[X_t \mid X_0 = x] = x \quad \text{for all } t \geq 0.
\]
Moran model via martingales.

First, we observe that the Moran process is a martingale: there

\[ p_j = q_j = \frac{j(n-j)}{n^2}, \quad r_j = 1 - 2\frac{j(n-j)}{n^2} \]

and

\[ E[X_{t+1} \mid X_t = j] = (j + 1) \cdot p_j + j \cdot r_j + (j - 1) \cdot q_j = j. \]

Moreover, the Optional Stopping Theorem will allow us to answer the question of finding the probability of fixation with all \( n \) alleles being \( A \),

\[ \alpha = P(X_{T_f} = n \mid X_0 = j). \]

Indeed, by the Optional Stopping Theorem,

\[ 0 \cdot (1 - \alpha) + n \cdot \alpha = E[X_{T_f} \mid X_0 = j] = j. \]

Thus,

\[ P(X_{T_f} = n \mid X_0 = j) = \alpha = \frac{j}{n}. \]
Wright-Fisher Model via martingales.

Given that we know $X_t$, variable $X_{t+1}$ is a Binomial random variables with $n = 2N$ trials and $p = \frac{X_t}{n}$, i.e.

$$P(X_{t+1} = k \mid X_t = j) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad \text{where } p = \frac{j}{n}.$$  

Then, $X_t$ is a martingale:

$$E[X_{t+1} \mid X_t = j] = np = j.$$  

Here too, by the Optional Stopping Theorem,

$$0 \cdot (1 - \alpha) + n \cdot \alpha = E[X_{T_f} \mid X_0 = j] = j$$  

and therefore

$$\alpha := P(X_{T_f} = n \mid X_0 = j) = \frac{j}{n}.$$  

Martingales.

**Definition.** A sequence of random variables $\{M_t\}$ is a martingale with respect to a homogeneous Markov chain $\{X_t\}$ if

- $M_t$ is a function of $X_t, X_{t-1}, \ldots, X_0$,
- $E[|M_t|] < \infty$ for all $t \geq 0$, and
- 

$$E[M_{t+1} \mid X_t, X_{t-1}, \ldots, X_0] = M_t.$$ 

**Optional Stopping Theorem.** Suppose $\{M_t\}$ is a martingale with respect to $\{X_t\}$, and $T$ is a stopping time with respect to $X_t$. If $P(T < \infty) = 1$ and there is $K > 0$ such that $|M_t| \leq K$ when $t < T$, then

$$E[M_T \mid X_0] = M_0.$$
Martingales and harmonic functions.

Suppose \( \{X_t\} \) is a time homogeneous Markov chain (HMC).

We say that \( h(\cdot) \) is a harmonic function with respect to the transition probabilities \( \{p(x,y)\} \) if \( h \) satisfies the averaging property

\[
\sum_y p(x,y)h(y) = h(x).
\]

There, \( h(X_t) \) is a martingale with respect to \( \{X_t\} \):

\[
E[h(X_{t+1}) \mid X_t = x] = \sum_y p(x,y)h(y) = h(x)
\]

and

\[
E[h(X_{t+1}) \mid X_t] = h(X_t).
\]
Martingales and harmonic functions.

For a birth-and-death chain $X_t$, the probability harmonic function $h$ is the one satisfying the averaging property

$$h(k) = q_k h(k - 1) + (1 - q_k - p_k) h(k) + p_k h(k + 1)$$

The above recurrence relation, after being simplified as

$$q_k \left( h(k) - h(k - 1) \right) = p_k \left( h(k + 1) - h(k) \right)$$

yields $h(0) = A$, $h(1) = A + B$, and

$$h(k) = A + B \left( 1 + \sum_{j=2}^{k} \frac{q_1 \cdots q_{j-1}}{p_1 \cdots p_{j-1}} \right) \quad \text{for } k = 2, 3, \ldots$$

Thus $M_t = h(X_t)$ is a martingale with respect to $\{X_t\}$. 
Martingales and harmonic functions.

For a birth-and-death chain $X_t$, the probability harmonic function $h$ is the one satisfying the averaging property

$$h(k) = q_k h(k-1) + (1 - q_k - p_k) h(k) + p_k h(k+1)$$

The above recurrence relation yields $h(0) = A$,

$$h(1) = A + B, \quad \text{and} \quad h(k) = A + B \left( 1 + \sum_{j=2}^{k} \frac{q_1 \cdots q_{j-1}}{p_1 \cdots p_{j-1}} \right) \quad \text{for } k = 2, 3, \ldots$$

Thus $M_t = h(X_t)$ is a martingale with respect to $\{X_t\}$. Define the following stopping time with respect to $X_t$,

$$T = \min\{t \geq 0 : X_t = 0 \text{ or } m\}.$$ 

Then, given that $X_0 = j$ for $0 \leq j \leq m$,

$$P(X_T = m \mid X_0 = j) = \frac{h(j) - h(0)}{h(m) - h(0)}.$$
Martingales and harmonic functions.

Example. For a birth-and-death chain $X_t$ with $p_k = p$ and $q_k = q$ for all $k$, and $p \neq q$,

$$h(k) = qh(k - 1) + (1 - q - p)h(k) + ph(k + 1)$$

yielding

$$h(k) = A + B \left( \frac{q}{p} \right)^k$$

for $k = 0, 1, 2, 3, \ldots$

Define the following stopping time with respect to $X_t$,

$$T = \min\{t \geq 0 : X_t = 0 \text{ or } m\}.$$

Then, given that $X_0 = j$ for $0 \leq j \leq m$,

$$P(X_T = m \mid X_0 = j) = \frac{h(j) - h(0)}{h(m) - h(0)} = \frac{\left( \frac{q}{p} \right)^j - 1}{\left( \frac{q}{p} \right)^m - 1}.$$