Re-Exam - Topics in Probability and Statistics (WMMA039-05) Solutions

Date and time: April 5, 2023, 11.45-13.45

Exercise 1

In this exercise we consider the Moran model in a popualtion of $2N$ haploid individuals, where N is large.

(a) Formulate the Moral model. $\boxed{10}$

(b) Consider two individuals currently living, let T be the time since the most recent common ancestor of the two individuals lived. Compute $\mathbb{E}[T]/N$. The answer may depend on details of your model formulation in part (a). $\boxed{10}$

Solution:

a) In a Moran model each living individual gives independently birth at rate 1 per time unit (this is the definition of the unit of time we use). At the moment an individual gives birth, one individual, uniformly chosen from the individuals that were already alive, dies.

b) Backward in time each lineage "dies" and continues on a uniformly chosen other lineage at rate $(2N-1)/(2N)$ because forward in time, the birthing individual can be the one that dies. So two given lineages coalesce at rate $[2/(2N-1)] \times [(2N-1)/(2N)] = 1/N$, because both lineages backward in time can die and then merge with the other lineage.

So $\mathbb{E}[T]$, the expected time until coalescence is N, and $\mathbb{E}[T]/N \to 1$ as $N \to \infty$.

Exercise 2

Consider the infinite sites variant of the standard Wright-Fisher model. Assume that recombination is not possible. Argue that for four individuals with a common ancestor it is impossible (within this model) to find a pair of sites on which the nucleotides are (A, A) for person 1, (A, C) for person 2, (C, A) for person 3 and (C, C) for person 4. $\boxed{15}$

Solution: In the infinite sites model we can have at most one mutation at a given site. The easiest way to see that we can have at most 3 combinations of nucleotides at any pair of two sites is that there are at most two mutations at the pair of sites and the total number of combinations of nucleotides we see in the population is at most 3 (the original plus two combinations caused by mutation).

Exercise 3

For this exercise recall the following. If we consider the genealogy of a sample of size n then the generations in this genealogy in which there are k lineages are called level k of the genealogy. If we put the k lineages at level k in random (uniform) order and consider the k-dimensional vector denoting how many of the sampled individuals descent from the k lineages separately, then this random vector is uniformly distributed on all k dimensional vectors with strictly positive integer vanlued entries, of which the entries sum to n.

In this exercise we consider the infinite sites variant of the Wright-Fisher model, The population is large and of size $2N$. Let μ be the mutation probability per generation per individual and let $\theta = 4N\mu$. Below assume that $2 \leq k \leq n$.

(a) Show that the probability that exactly one individual in the sample of size n is a descendant of a given uniformly chosen lineage at level k is given by

$$\frac{k-1}{n-1}.$$

$\boxed{5}$

(b) Compute the expected number of mutations that are carried by only 1 of the n individuals in the sample. $\boxed{15}$

Solution:

a) See page 54 of course book and note that for $k > 2$

$$\frac{k-1}{n-1} = \binom{n-1-1}{k-2} \Big/ \binom{n-1}{k-1}$$

b) Using part a), we look at the expected total length of the ancestral lineages at each level from 2 to n, which are ancestral to one individual at current level. At level k there are k lineages each of which has expected length $1/\binom{k}{2}$ and probability $\frac{(k-1}{n-1}$ to be ancestor of exactly one individual living now.

So the expected total length of the ancestral lineages at level k that are ancestral to 1 current individual is

$$k \times \frac{1}{\binom{k}{2}} \times \frac{(k-1)}{(n-1)} = \frac{2}{n-1}$$

So, we are interested in

$$\sum_{k=2}^{n} \frac{2}{(n-1)} = 2.$$

The rate at which mutations occur at linesegment is $\theta/2$ in rescaled time. So expected number of unique mutations is θ.

Exercise 4

In this exercise we consider the Wright-Fisher model in a growing popualtion. Assume that one generation corresponds with one time unit. Let N_0 be a very large integer and $\rho > 0$ a constant. The population consists currently (say at time $t = 0$) of $2N(0) = 2N_0$ haploid individuals and the population size at time $s \leq 0$ is well approximated by $2N(s)$, where $N(s) = N_0 e^{\rho s/N_0}$. Note that $N(s) < N(0)$, for $s < 0$. Suppose that we have a sample of size 2.

Let T_2 be the last time there was a common ancestor of the 2 individuals in the sample (i.e. T_2 is minus the first coalescence time in the corresponding colalescence model).

For fixed $t > 0$ and $N_0 \to \infty$ compute

$$\mathbb{P}(-T_2/N_0 > t).$$

Explicitly compute the (asymptotic) probability that this most recent common ancestor lived in a population of size at most N_0. $\boxed{20}$

Solution: See page 431 of course book equation 4.4 Probability of no coalescence in past t generations is

$$\prod_{s=-t}^{-1}\left(1 - \frac{1}{2N(s)}\right) \approx exp\left(-\sum_{s=-t}^{-1}\frac{1}{2N(s)}\right) \approx exp\left(-\int_{-t}^{0}\frac{e^{-\rho s/N_0}}{2N_0}ds\right) = exp\left(-\frac{e^{\rho t/N_0} - 1}{2\rho}\right).$$

So the probability no coalescence in the past tN_0 generations (which is what the question was) is

$$exp\left(-\frac{e^{\rho t} - 1}{2\rho}\right).$$

The population was size N_0 at the time when $e^{\rho s/N_0} = 1/2$. Filling that in (considering that s is negative), we obtain that the probability that this most recent common ancestor lived in a population of size at most N_0 is given by $exp\left(-\frac{2-1}{2\rho}\right) = e^{-1/(2\rho)}$.

Exercise 5

In this exercise we consider the Wright-Fisher model with two loci (say the a and b locus) on the same chromosome, where recombination is possible: Each individual independently copies both loci from one uniform individual in the previous generation with probability $1-r$, or with probability r it copies the two loci from two independent uniformly chosen individuals from the previous generation. As usual the population consists of $2N$ haploid individuals. We set $r = \rho/(4N)$, where ρ is a strictly positive constant. Assume that $2N$ generations occur in one time unit and consider the process in the large population limit.

Suppose that we start with a sample of two individuals. What is the expected time since both the genetic material on the a locus and on the b locus in the sampled individuals have a common ancestor?

Hint: We may denote the states that the ancestry may be in, by (x_1, x_2, x_3), where

x_1 is the number of individuals in a generation that is both an ancestor of an a locus in the sample and not of any b locus in the sample.

x_2 is the number of individuals in a generation that is both an ancestor of an b locus in the sample and not of any a locus in the sample.

x_3 is the number of individuals in a generation that is both an ancestor of an a locus in the sample and of a b locus in the sample.

The current state is $(0,0,2)$, and T is the last time the process was in either state $(0,0,1)$ or $(1,1,0)$. We are interested in $-\mathbb{E}[T]$.

It turns out that you can simplify the state space by looking at the evolution (backwards in time) of the geneology over the states $A = (0,0,2)$, $B = (1,1,1)$, $C = (2,2,0)$, $D = (1,0,1) \cup (0,1,1) \cup (2,1,0) \cup (1,2,0)$ and $E = (0,0,1) \cup (1,1,0)$. Furthermore, you can then describe the "ancestries" of the two loci back in time as a Markov Process on those states and you are interested in the expected time it takes to go from state A to state E.

$\boxed{15}$

Solution: Use ideas of page 83/85 of course book. Let T_X be the expected time it takes to move from state X to to state E (possibly via other states), where $X \in \{A, B, C, D, E\}$. Clearly $T_E = 0$ and we are interested in T_A. Let J_X ($X \in \{A, B, C, D, E\}$) be the time it takes to jump from state X to any other state.

From state A the jump rates are 1 to state E (the linages coalesce) and $2\rho/2 = \rho$ to go to state B (one of the two "coupled" pairs split; each pair splits at rate $2N\rho/(4N) = \rho/2$). So the total rate of leaving state A is $\rho + 1$, that is $J_A = 1/(\rho+1)$. This leads to

$$T_A = J_A + \frac{1}{\rho+1}T_E + \frac{\rho}{\rho+1}T_B = \frac{1}{\rho+1} + \frac{1}{\rho+1}T_E + \frac{\rho}{\rho+1}T_B,$$

which implies (using $T_E = 0$)

$$(\rho+1)T_A = 1 + \rho T_B. \tag{1}$$

Similarly, the rate of going from state B to state A is 1 (when the two "non-paired" lineages get paired again), the rate of going from state B to state C is $\rho/2$ (when the paired loci split) and rate of going from state B to state D is 2 (when either one of the non-paired lineages coalesces with the paired one). So, the total rate of leaving state B is $\rho/2+3$. This leads to

$$T_B = J_B + \frac{1}{\rho/2+3}T_A + \frac{\rho/2}{\rho/2+3}T_C + \frac{2}{\rho/2+3}T_D = \frac{1}{\rho/2+3} + \frac{1}{\rho/2+3}T_A + \frac{\rho/2}{\rho/2+3}T_C + \frac{2}{\rho/2+3}T_D,$$

which implies

$$(\rho/2+3)T_B = 1 + T_A + (\rho/2)T_C + 2T_D. \qquad (2)$$

Similarly the rate of going from state C to state B is 4 (there are 4 possible pairs of site 1 and site 2 that can coalesce) while the rate of going from state C to state D is 2 (each of the sites can coalesce). So,

$$T_C = \frac{1}{6} + \frac{4}{6}T_B + \frac{2}{6}T_D. \qquad (3)$$

Finally from stat D we can only go to state E, which occurs if on the site where still 2 lineages are present coalescence occurs. This occurs at rate 1. So

$$T_D = 1 + T_E = 1.$$

Filling in $T_D = 1$ in (3) gives

$$6T_C = 3 + 4T_B \qquad (4)$$

Filling in $T_D = 1$ and (4) in (2) give

$$(\rho/2+3)T_B = 3 + T_A + (\rho/12)(3+4T_B) \Rightarrow (2\rho+36)T_B = (3\rho+36) + 12T_A. \qquad (5)$$

Filling in (5) into (1) then gives.

$$(\rho+1)T_A = 1 + \rho\left(\frac{3\rho+36}{2\rho+36} + \frac{12}{2\rho+36}T_A\right).$$

or

$$2(\rho^2 + 13\rho + 18)T_A = (3\rho^2 + 38\rho + 36).$$

That is,

$$T_A = \frac{(3\rho^2 + 38\rho + 36)}{2(\rho^2 + 13\rho + 18)}.$$