CS 3824
Solutions to Homework on Probability

Lenwood S. Heath

February 13, 2008

[30] 1. Let X_1 and X_2 be independent random variables, where X_1 has an exponential distribution with parameter α_1 and X_2 has an exponential distribution with parameter α_2. Let $Z = \min\{X_1, X_2\}$ be the random variable that is the minimum of X_1 and X_2. (If X_1 and X_2 represent the times at which two different mutations in a genome occur, then Z is the time to the first mutation.)

A. Compute $F_Z(t)$ as a function of $F_{X_1}(t)$ and $F_{X_2}(t)$.

B. Derive $f_Z(t)$.

C. What is the distribution of Z?

D. Generalize the above in the following context. Consider 100 consecutive positions on a chromosome. Assume that these positions undergo point mutations independently, each at a rate $\alpha = 1/(10^5$ years). What is the expected amount of time to the first mutation?

A. Start with the definitions of these cumulative distribution functions:

$$\begin{aligned} F_{X_1}(t) &= \Pr[X_1 \le t] \\ F_{X_2}(t) &= \Pr[X_2 \le t] \\ F_Z(t) &= \Pr[Z \le t]. \end{aligned}$$

Now, use these definitions and the rules of probability to obtain the desired result:

$$\begin{aligned} F_Z(t) &= \Pr[Z \le t] \\ &= 1 - \Pr[Z > t] \\ &= 1 - \Pr[\min\{X_1, X_2\} > t] \\ &= 1 - \Pr[(X_1 > t) \wedge (X_2 > t)] \\ &= 1 - \Pr[X_1 > t] \Pr[X_2 > t] \\ &= 1 - (1 - F_{X_1}(t))(1 - F_{X_2}(t)) \\ &= F_{X_1}(t) + F_{X_2}(t) - F_{X_1}(t) F_{X_2}(t). \end{aligned}$$

Here, we used the independence of X_1 and X_2 to obtain $\Pr[(X_1 > t) \wedge (X_2 > t)] = \Pr[X_1 > t] \Pr[X_2 > t]$.

B. First, we expand the expression we obtained for $F_Z(t)$:

$$\begin{aligned} F_Z(t) &= F_{X_1}(t) + F_{X_2}(t) - F_{X_1}(t) F_{X_2}(t). \\ &= 1 - e^{-\alpha_1 t} + 1 - e^{-\alpha_2 t} - \left(1 - e^{-\alpha_1 t}\right)\left(1 - e^{-\alpha_2 t}\right) \\ &= 1 - e^{-(\alpha_1 + \alpha_2) t}. \end{aligned}$$

Now, we compute as follows:

$$\begin{aligned} f_Z(t) &= \frac{d}{dt} F_Z(t) \\ &= \frac{d}{dt}\left(1 - e^{-(\alpha_1+\alpha_2)t}\right) \\ &= (\alpha_1 + \alpha_2) e^{-(\alpha_1+\alpha_2)t}. \end{aligned}$$

C. From the density function $f_Z(t)$, we see that Z has an exponential distribution with parameter $\alpha_1 + \alpha_2$.

D. Generalizing the previous observation, we get that the minimum of any set of independent, exponentially distributed random variables is an exponentially distributed random variable with parameter the sum of the parameters. If there are 100 such random variables, each with parameter α, then the minimum has parameter 100α and expected value $1/(100\alpha)$. For the particular value of α, we get the expected value $1/(100(1/10^5)) = 10^3$ years.

---

[30] 2. Let S be the following nucleotide sequence:

AATCTGTACGGTGCTAGTTACCTGCACAGTTCGGACCTGCCAATGCCGTAAGC

For simplicity, assume that S is circular — the C at the end is followed by the A at the beginning.

A. Use the sequence S to construct a 4-state Markov chain M representing the character transition frequencies of S.

B. Compute the stationary distribution for M.

C. If M is in state C, what is the probability that the next 10 characters are

ACCGAACGAG?

D. If M is in state C, determine the most likely sequence of 10 characters to come next. Compare the probability of those 10 characters to the probability from Part **C**. What tentative conclusion do you reach?

A. Start with a count of each nucleotide:

Letter	Count
A	12
C	15
G	13
T	13

Now, count all pairs of nucleotides, including the CA implied by S being circular:

	A	C	G	T
A	3	4	3	2
C	4	4	3	4
G	1	5	2	5
T	4	2	5	2

The transition probability matrix for M is gotten by dividing the entries in each row by the letter frequency of the nucleotide for that row. Hence,

$$P = \begin{pmatrix} 1/4 & 1/3 & 1/4 & 1/6 \\ 4/15 & 4/15 & 1/5 & 4/15 \\ 1/13 & 5/13 & 2/13 & 5/13 \\ 4/13 & 2/13 & 5/13 & 2/13 \end{pmatrix}$$

B. The stationary distribution is the matrix

$$\Pi = \begin{pmatrix} \pi_{\mathrm{A}} & \pi_{\mathrm{C}} & \pi_{\mathrm{G}} & \pi_{\mathrm{T}} \end{pmatrix}$$

such that $\pi_{\mathrm{A}}+\pi_{\mathrm{C}}+\pi_{\mathrm{G}}+\pi_{\mathrm{T}} = 1$ and $\Pi = \Pi P$. We get the following system of equations:

$$\begin{aligned} \pi_{\mathrm{A}} + \pi_{\mathrm{C}} + \pi_{\mathrm{G}} + \pi_{\mathrm{T}} &= 1 \\ \frac{1}{4}\pi_{\mathrm{A}} + \frac{4}{15}\pi_{\mathrm{C}} + \frac{1}{13}\pi_{\mathrm{G}} + \frac{4}{13}\pi_{\mathrm{T}} &= \pi_{\mathrm{A}} \\ \frac{1}{3}\pi_{\mathrm{A}} + \frac{4}{15}\pi_{\mathrm{C}} + \frac{5}{13}\pi_{\mathrm{G}} + \frac{2}{13}\pi_{\mathrm{T}} &= \pi_{\mathrm{C}} \\ \frac{1}{4}\pi_{\mathrm{A}} + \frac{1}{5}\pi_{\mathrm{C}} + \frac{2}{13}\pi_{\mathrm{G}} + \frac{5}{13}\pi_{\mathrm{T}} &= \pi_{\mathrm{G}} \\ \frac{1}{6}\pi_{\mathrm{A}} + \frac{4}{15}\pi_{\mathrm{C}} + \frac{5}{13}\pi_{\mathrm{G}} + \frac{2}{13}\pi_{\mathrm{T}} &= \pi_{\mathrm{T}} \end{aligned}$$

Solving this system of equations, we get the unique solution

$$\begin{pmatrix} \pi_{\mathrm{A}} & \pi_{\mathrm{C}} & \pi_{\mathrm{G}} & \pi_{\mathrm{T}} \end{pmatrix} = \begin{pmatrix} 12/53 & 15/53 & 13/53 & 13/53 \end{pmatrix}.$$

C. The desired probability is the product of 10 entries from P, as follows:

$$\begin{aligned} P_{\mathrm{CA}}P_{\mathrm{AC}}P_{\mathrm{CC}}P_{\mathrm{CG}}P_{\mathrm{GA}}P_{\mathrm{AA}}P_{\mathrm{AC}}P_{\mathrm{CG}}P_{\mathrm{GA}}P_{\mathrm{AG}} &= \frac{4}{15}\cdot\frac{1}{3}\cdot\frac{4}{15}\cdot\frac{1}{5}\cdot\frac{1}{13}\cdot\frac{1}{4}\cdot\frac{1}{3}\cdot\frac{1}{5}\cdot\frac{1}{13}\cdot\frac{1}{4} \\ &= \frac{1}{8555625} \\ &\approx 1.16882\times 10^{-7}. \end{aligned}$$

D. Finding the most likely sequence of 10 characters is a more difficult problem than I intended. Instead of an optimal solution, I will present the greedy solution.

In the greedy solution, one always takes the highest probability transition (HPT) out of each state. Here is a table of the HPTs:

State	HPT
A	G
C	T
G	C
T	G

In the case of ties, an arbitrary HGT has been chosen. We can now read the greedy sequence by following the table:

TGCTGCTGCT

The corresponding probability of the sequence is the product

$$\begin{aligned} P_{\mathrm{CT}}P_{\mathrm{TG}}P_{\mathrm{GC}}P_{\mathrm{CT}}P_{\mathrm{TG}}P_{\mathrm{GC}}P_{\mathrm{CT}}P_{\mathrm{TG}}P_{\mathrm{GC}}P_{\mathrm{CT}} &= P_{\mathrm{CT}}^4 P_{\mathrm{TG}}^3 P_{\mathrm{GC}}^3 \\ &= \left(\frac{4}{15}\right)^4 \cdot \left(\frac{5}{13}\right)^3 \cdot \left(\frac{5}{13}\right)^3 \\ &= \frac{6400}{390971529} \\ &\approx 2.63695 \times 10^{-5}. \end{aligned}$$

The ratio of this probability to the previous probability is

$$\begin{aligned} \frac{\dfrac{6400}{390971529}}{\dfrac{1}{8555625}} &= \frac{4000000}{28561} \\ &\approx 140.051. \end{aligned}$$

The high probability sequence is about 140 times more likely than the example sequence. We conclude that the example sequence is an unlikely sequence for this Markov chain to generate after C.