MATH-578B: Assignment # 3

Due on Thursday, October 22, 2015

Saket Choudhary
2170058637

Contents

Problem 1 3

Problem 2 4

Problem 3 5

Problem 4 7

Exercise 11.9 8

Exercise 11.11 8

Problem 1

The coverage c depends on the position x as: $c = \frac{NL_x}{G}$ where L_x is the expected length of of clones covering x.
Probability any position x to be covered by atleast one clone = (1− Probability that it is sequenced by atleast one clone).
Probability that position x is not sequenced = Probability of zero clones starting in $(x-L, x]$ = No arrivals in the interval $(x-L, x] = e^{-c(x)}$
Probability that it is sequenced $= 1 - e^{-c(x)}$ where $c(x)$ represents that c is a function of x.
$C \sim \Gamma(\alpha, \beta)$

$$f(c) = \frac{c^{\alpha-1}e^{-c/\beta}}{\beta^{\alpha}\Gamma(\alpha)}$$

Thus,

$$P(N_h = k) = \int_0^{\infty} e^{-ch}\frac{(ch)^k}{k!} \times \frac{c^{\alpha-1}e^{-c/\beta}}{\beta^{\alpha}\Gamma(\alpha)}dc$$

Problem 2

Given: $\lim_{n\to\infty}(1 - F(b\log(n) + x/a)) = G(x)$

$$\lim_{n\to\infty}(1 - F(b\log(n) + x/a)) = G(x)$$
$$\lim_{n\to\infty} F(b\log(n) + x/a) = 1 - G(x)/n$$

$$\begin{aligned}
P(a(max_i X_i - b\log(n)) \leq x) &= P(max_i X_i \leq x/a + b\log(n)) \\
&= P(X_1 \leq x/a + b\log(n))P(X_2 \leq x/a + b\log(n)) \ldots P(X_n \leq x/a + b\log(n)) \\
&= (F(x/a + b\log(n)))^n \\
&= \lim n_i nfty(1 - G(x)/n)^n \\
&= \lim_{n\to\infty} e^{n\log(1-G(x)/n)} \\
&= e^{-G(x)}
\end{aligned}$$

Choosing a, b for $G(x) = e^{-x}$ given $X_i \sim exponential(\lambda)$
$f(x|\lambda) = \lambda e^- \lambda x \implies F(x) = 1 - e^{-\lambda x}$
Now,

$$\begin{aligned}
\lim n \to \infty 1 - G(x)/n &= F(b\log(n) + x/a) \\
&= 1 - e^{-\lambda(b\log(n)+x/a)} \\
e^{-x}/n &= e^{-\lambda(b\log(n)+x/a)} \\
-x &= \log(n) + -\lambda(b\log(n) + x/a) \\
x(-1 + \lambda/a) &= \log(n) - b\lambda\log(n)
\end{aligned}$$

Thus,

$$a = \lambda; b = \frac{1}{\lambda}$$

Problem 3

Target Distribution in aligned region: $P(R,R) = 0.2$; $P(Y,Y) = 0.7$; $P(R,Y) = 0.1$

$$\xi_r = 0.2$$
$$\xi_y = 0.8$$

By Theorem 11.7 we have that thte proportion of letter a aligning with letter b in the best matching interval converges to:

$$p(a,b) = \xi_a \xi_b p^{-s(a,b)}$$

Equivalently:

$$s(a,b) = \log_{1/p}(\frac{p(a,b)}{\xi_a \xi_b})$$

$$p = \xi_r \xi_r + \xi_y \xi_y = 0.68$$

Thus

$$\begin{aligned} P(RR) &= \xi_r \xi_r p^{-s(r,r)} \\ s(r,r) &= \log_{1/0.68}(\frac{0.2}{0.04}) \\ &= 4.17 \end{aligned}$$

$$\begin{aligned} s(r,y) &= \log_{1/p}(\frac{p(r,y)}{\xi_r \xi_y}) \\ s(r,y) &= \log_{1/0.68}(\frac{0.1}{0.16}) \\ &= -1.21 \end{aligned}$$

$$\begin{aligned} s(y,y) &= \log_{1/p}(\frac{p(y,y)}{\xi_y \xi_y}) \\ s(y,y) &= \log_{1/0.68}(\frac{0.7}{0.64}) \\ &= 0.23 \end{aligned}$$

$$\boxed{s(r,r) = 4.17}$$
$$\boxed{s(y,y) = 0.23}$$
$$\boxed{s(y,r) = -1.21}$$

Problem 3 continued on next page. . .

To find the value of λ such that:

$$\lim_{n\to\infty, m\to\infty} P\{\lambda R_{mn} - \log(K_{mn}) < x\} = exp(-exp(-x))$$

$$\lambda = log(1/p) = 0.38$$

And given that the score for 1000bp alignment is 100, the p value is given by;

$$p - value = 1 - e^{-e^{-s}}$$

where $s = \lambda R_{mn} - \log(Kmn)$
If the p-value is less than some pre-defined threshold, the hypothesis that alignment is as good as by random chance can be rejected.

Problem 4

Minimal neighborhood set $J_{i,j}$ such that $\{i', j' \in J^c_{i',j'}\}$ are independent of $Y_{i,j}$ is given by: $\{(i', j') : |i - i'| \leq tor|j - j'| \leq t\}$
Now,

$$\begin{aligned}
b_1 &= \sum_{i \in I} \sum_{j in J_i} E(X_i)E(X_j) \\
&= p^t \sum_{j \in J_i} E(X_j) + \sum_{i=2}^{n-t+1} (1-p)p^t \sum_{j \in J_i} E(X_j) \\
&= (n-t+1)p^t(2t+1)p^t \times 2 + (n-t+1)^2(1-p)^2 p^{2t}(4t+2) \\
&= p^{2t}(n-t+1)(4t+2)(1+(n-t+1)(1-p)^2)
\end{aligned}$$

$$\begin{aligned}
E[NC_n] &= (n-t+1)^2(1-p)p^t \\
&= \lambda
\end{aligned}$$

Thus,

$$n^2(1-p)p^t \sim \lambda$$

$$\begin{aligned}
\log(n^2(1-p)) + t\log(p) &= \log(\lambda) \\
t &= -\log_{1/p}(\lambda) + 2\log_{1/p}(n(1-p))
\end{aligned}$$

And $b1$ can be approximated as:

$$(n-t+1)^2(4t+2)(1-p)^2 p^{2t}$$

If we approximate $t_n = 2\log_{1/p} n + x$

$$\begin{aligned}
(n-t+1)^2(4t+2)(1-p)^2p^{2t} &= (n - 2\log_{1/p} n + 1)^2(4(2\log_{1/p} n + x) + 2)((1-p)p^{2\log_{1/p} n + x})^2 \\
&= (n - 2\log_{1/p} n + 1)^2(4(2\log_{1/p} n + x) + 2)((1-p) \times 1/n^2)^2 \to 0 \\
&\to 0
\end{aligned}$$

Exercise 11.9

Part (a): $\xi_a = 1/|A|$

$$p = \sum_{a \in A} \xi_a \xi_a = |A|/|A|^2 = 1/|A|$$

$$\begin{aligned}\mu_{a,a} &= \xi_a \xi_a p^{-s(a,a)} \\ &= 1/|A|^2 \times |A| \\ &= 1/|A|\end{aligned}$$

Part (b): To Derive $s(a,b)$ such that $\mu_{a,a} = 1/|A|$ We have(from problem 3):

$$s(a,b) = \log_{1/p}(\frac{p(a,b)}{\xi_a \xi_b})$$

And hence:

$$s(a,a) = \log_{|A|}(\frac{1/|A|}{1/|A|^2}) = 1$$

Thus,

$$s(a,a) = 1; s(a,b) = -\infty$$

Exercise 11.11

$E[NC] = (n-t+1)^N p^t$
The probability of x being the starting position of alignment in the N sequences: $p = (\frac{1}{|A|})^N \times |A| = 1/|A|^{N-1} = 1/n^{N-1}$ (since A,B are iid)

$$|A| = n$$

Thus, assuming the longest run is unique, we have that the expected number of runs of this length be 1:

$$\begin{aligned}(n-t+1)^N p^t &= 1 \\ (n-t+1)^N 1/n^{Nt-t} &= 1 \\ n^{t-Nt+N} &\approx 1 \\ t &= \frac{1}{N-1} \log_{1/n} N\end{aligned}$$

And hence as

$$N \to \infty \implies t \to 1/N^2 \to 0$$