MATH-578B: Midterm

Due on Thursday, November 5, 2015

Saket Choudhary
2170058637

Contents

Problem 1

Problem 1: (a)

$$P = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix}$$

Let the stationary state be given by $\pi = (\pi_1, \pi_2)$, then:

$$\pi.P = \pi$$
$$\pi_1 + \pi_2 = 1$$

Solving which gives:

$$(1-\alpha)\pi_1 + \pi_2 = 1$$
$$\pi_1 + \pi_2 = 1$$
$$\implies (\pi_1, \pi_2) = (\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta})$$

Problem 1: (b)

$\mathbf{w} = 101$

$$\beta_{w,w}(0) = 1$$
$$\beta_{w,w}(1) = 0$$
$$\beta_{w,w}(2) = 1$$

$$P_u(0) = 1$$
$$P_u(1) = p_{w_2w_3} = p_{01} = \alpha$$
$$P_U(2) = p_{w_1w_2}p_{w_2w_3} = p_{10}p_{01} = \beta\alpha$$

Now,

$$G_{w,w}(t) = \sum_{j=0}^{2} t^j \beta_{w,w}(j) P_{w,w}(j)$$

Thus,

$$\begin{aligned} G_{w,w}(t) &= 1 \times 1 \times 1 + t \times 0 \times \alpha + t^2 \times 1 \times \beta\alpha \\ &= 1 + \alpha\beta t^2 \end{aligned}$$

Problem 1 continued on next page. . .

Problem 1: (c)

X_n : Number of occurrences(overlaps allowed) in $A_1A_2A_3\dots A_n$ Using Theorem 12.1:

$$\lim_{n\to\infty}\frac{1}{n}E(X_n) = \pi_w = \pi_1 \times p_{10} \times p_{01} = \frac{\beta}{\alpha+\beta} \times \beta \times \alpha = \frac{\alpha\beta^2}{\alpha+\beta}$$

Thus

$$\lim_{n\to\infty}\frac{X_n}{n} = \frac{\alpha\beta^2}{\alpha+\beta}$$

Problem 1: (d)

Spectral decomposition of P:

$$det \begin{bmatrix} \alpha-\lambda & 1-\alpha \\ 1-\beta & \beta-\lambda \end{bmatrix} = 0$$

$$\lambda^2 + (\alpha+\beta-2)\lambda + (1-\alpha-\beta) = 0$$

Thus, $\lambda_1 = 1$ and $\lambda_2 = 1-\alpha-\beta$
Eigenvectors are given by:
$v_1^T = \begin{pmatrix} x_1 & x_1 \end{pmatrix} \; \forall \; x_1 \in R$
and for λ_2 , $v_2 = \begin{pmatrix} x_1 & \frac{-\beta x_1}{\alpha} \end{pmatrix}$
Now using Markov property: $P(X_n = 1 | X_0 = 0) = (P^n)_{01}$
Now,
$P^n = VD^nV^{-1}$
where:

$$V = \begin{bmatrix} 1 & 1 \\ 1 & \frac{-\beta}{\alpha} \end{bmatrix}$$

and

$$D = \begin{bmatrix} 1 & 0 \\ 0 & (1-\alpha-\beta) \end{bmatrix}$$

$$V^{-1} = \frac{-1}{\frac{\beta}{\alpha}+1} \begin{bmatrix} -\frac{\beta}{\alpha} & -1 \\ -1 & 1 \end{bmatrix}$$

Thus,

$$P^n = \begin{bmatrix} 1 & 1 \\ 1 & \frac{-\beta}{\alpha} \end{bmatrix} \times \begin{bmatrix} 1 & 0 \\ 0 & (1-\alpha-\beta)^n \end{bmatrix} \times \frac{-1}{\frac{\beta}{\alpha}+1} \begin{bmatrix} -\frac{\beta}{\alpha} & -1 \\ -1 & 1 \end{bmatrix}$$

$$P^n = \frac{1}{\alpha+\beta} \begin{bmatrix} \beta + \alpha(1-\alpha-\beta)^n & \alpha - \alpha(1-\alpha-\beta)^n \\ \beta - \beta(1-\alpha-\beta)^n & \alpha + \beta(1-\alpha-\beta)^n \end{bmatrix}$$

ALITER
We consider the following identity: $P^{n+1} = PP^n$
then:

$$\begin{bmatrix} p_{00}^{n+1} & p_{01}^{n+1} \\ p_{10}^{n+1} & p_{11}^{n+1} \end{bmatrix} = \begin{bmatrix} p_{00}^{n} & p_{01}^{n} \\ p_{10}^{n} & p_{11}^{n} \end{bmatrix} \times \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix}$$

$\implies$

$$\begin{aligned} p_{11}^{n+1} &= p_{10}^n(\alpha) + p_{11}^n(1-\beta) \\ &= (1-p_{11}^n)(\alpha) + (p_{11}^n)(1-\beta) \\ &= \alpha + (1-\alpha-\beta)p_{11}^n \end{aligned} \tag{1d.1}$$

On similar lines:

$$p_{00}^{n+1} = (1-\alpha-\beta)p_{00}^{n} + \beta \tag{1d.2}$$

In order to solve equations of type $1d.1$ and $1d.2$ we take the following strategy:
By substituting $p_{00}^{n+1} = p_{00}^{n}$ (and thus obtaining the stationary solution at $\frac{\beta}{\alpha+\beta}$), $1d.2$ can be reduced to the following form:

$$p_{00}^{n+1} = \frac{\beta}{\alpha+\beta} = (1-\alpha-\beta)(p_{00}^{n} - \frac{\beta}{\alpha+\beta})$$

Let's call $y^(n) = p_{00}^{n} - \frac{\beta}{\alpha+\beta}$
Then $1d.3$ is similar to:

$$y^{(n+1)} = (1-\alpha-\beta)y^(n)$$
$$y^{(n+1)} = (1-\alpha-\beta)^n y^(0)$$

$y^(0) = p_{00}^{(0)} - \frac{\beta}{\alpha+\beta}$
Assume $p_{00}^{(0)} = 1 \implies y^{(0)} = \frac{\alpha}{\alpha+\beta}$
Which gives:

$$p_{00}^{n} = (1-\alpha-\beta)^n \frac{\alpha}{\alpha+\beta} + \frac{\beta}{\alpha+\beta}\text{if } \alpha+\beta > 0$$

Similarly,

$$p_{11}^{n} = (1-\alpha-\beta)^n \frac{\beta}{\alpha+\beta} + \frac{\alpha}{\alpha+\beta}\text{if } \alpha+\beta > 0$$

NOTE: If $\alpha+\beta = 0$, we get:

$$P^n = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Problem 1: (e)

$\pi_w = \pi_i p_{10} p_{01} = \frac{\alpha}{\alpha+\beta} \times \beta\alpha = \frac{\alpha^2\beta}{\alpha+\beta}$

$$\begin{aligned}
\lim_{n\to\infty} \frac{Var(X_n)}{n} &= 2\pi_w\big((\beta_{w,w}(0)P_w(0) - \pi_w) + (\beta_{w,w}(1)P_w(1) - \pi_w) + (\beta_{w,w}(2)P_w(2) - \pi_w)\big) \\
&+ 2\pi_w P_w(2)\sum_{j=0}^{\infty}\{p_{11}^{j+1} - \pi_1\} + \pi_w^2 - \pi_w \\
&= 2\frac{\alpha^2\beta}{\alpha+\beta}(1 + \alpha\beta - 3\frac{\alpha^2\beta}{\alpha+\beta}) + 2\frac{\alpha^2\beta}{\alpha+\beta}\sum_{j=0}^{\infty}\{\frac{\beta}{\alpha+\beta}(1-\alpha-\beta)^j\} - \frac{\alpha^2\beta}{\alpha+\beta} + (\frac{\alpha^2\beta}{\alpha+\beta})^2 \\
&= \frac{\alpha^2\beta}{\alpha+\beta}(\frac{(2\alpha + 2\beta - 4\alpha^2\beta + 2\alpha\beta^2 + 2\alpha\beta + 2\beta^2 + \alpha^2\beta - \alpha - \beta)}{\alpha+\beta}) \\
&= \frac{\alpha^2\beta(\alpha + \beta - 3\alpha^2\beta + 2\alpha\beta^2 + 2\alpha\beta + 2\beta^2)}{(\alpha+\beta)^2}
\end{aligned}$$

Problem 1: (f)

Y_n: Number of occurrences of word w in all words, thus, $Y_n \approx cX_n$
NOTE: I assume the problem should be to estimate $\lim_{n\to\infty} \frac{Y_n}{n}$ and $\lim_{n\to\infty} \frac{Var(Y_n)}{n}$ In the problem it mentions X_n instead of Y_n
Thus,

$$
\begin{aligned}
\lim_{n\to\infty} \frac{Y_n}{n} &= c \times \lim_{n\to} \frac{X_n}{n} \\
&= c\frac{\alpha^2\beta}{\alpha+\beta} \\
\lim_{n\to\infty} \frac{Var(Y_n)}{n} &= c^2 \times \lim_{n\to\infty} \frac{Var(X_n)}{n} \\
&= c^2 \times (\frac{\alpha^2\beta(\alpha+\beta-3\alpha^2\beta+2\alpha\beta^2+2\alpha\beta+2\beta^2)}{(\alpha+\beta)^2})
\end{aligned}
$$

Please see Appendix 1 for code and report.

Problem 2

Problem 2: (a)

Expected number of squares of side length t such that all X_v are 1 in the square: We simply choose a position on the positive lattice x axis and then construct a square around it $[(x_0, y_0)(x_0+t, y_0+t)]$ so we have $n-t+1$ choices for x_0 and $n-t+1$ choices for y_0, the constraint being that all points inside are all 1, there are approximately t^2 integer points inside

$E(\#\text{number of squares of side length } t \text{ such that all } X_v \text{ are 1 inside}) = (n-t)\times(n-t)p^{t^2} = (n-t+1)^2p^{t^2}$

More formally, we define the indicator $C_{i,j} = I(X_{i+p,j+q} = 1)\ \forall 0 \le p, q \le t-1$
and hence $E(Y_t) = E(\sum_{i=0}^{n-t+1}\sum_{j=0}^{n-t+1} C_{i,j}(t)) = (n-t+1)^2p^{t^2}$

Problem 2: (b)

Just like the largest run problem, this problem should satisfy

$$
\begin{aligned}
(n-t)^2 \times p^{T^2} &= 1 \\
\implies \lim_{n\to\infty} (n^2)p^{T_n^2} &= \frac{1}{n^2} \\
\lim_{n\to\infty} \log_{1/p} p^{T^2} &= \log_{1/p}(\frac{1}{n^2}) \\
\lim_{n\to\infty} T_n^2 &= 2\log_{1/p}(n) \\
\lim_{n\to\infty} \frac{T_n}{\sqrt{2\log_{1/p}(n)}} &= 1
\end{aligned}
$$

Thus,

$$a(n) = \sqrt{2\log_{1/p}(n)}$$

Problem 2: (c)

For declumping:

$$W_(1,1) = 1$$
$$W_{i,j} = (1 - \prod_{p=0} I(X_{i+p,j-1} = 1)I(X_{i-1,j+p} = 1)) \times \prod_{p=0}^{t-1}\prod_{q=0}^{t-1} I(X_{i+p,j+q} = 1)$$
$$W_(i,j) = C_{i,j}(t-1) - C_{i,j}(t)$$

The set I is given by: $I = \{(i,j) : 0 \le i,j \le n-t+1\}$ The dependence set for $\nu = (i,j)$ is given by $J_\nu = \{(i',j') \in I : |i-i'| \le t\text{and } |j-j'| \le t\}$
Now,

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

And,

$$\begin{aligned} b_2 &= \sum_{i\in I}\sum_{i\neq j\in J_i} E(X_iX_j) \\ &= 0 \end{aligned}$$

In order to choose a $t(n)$ such that W is approximately poisson with $\lambda = (n-t+1)^2(p^{(t-1)^2} - p^{t^2})$, choose $t_n = \sqrt{(2\log_{1/p}(n))}$ so that $b1 \to 0$ (because $b1 = \frac{(2t+1)^2\lambda^2}{(n-t+1)^2} \to \frac{\log(n)}{n^2} \to 0$) and hence by Thorem 11.22, we have W to be a poisson(since $b_1 = b_2 = 0$)

Problem 3

Problem 3: (a)

$$P(A_i = B_i) = \sum_{a\in S} \eta_a\gamma_a$$

Problem 3: (b)

If X_i is the the number of matches between two consecutive matches, it follows a **negative binomial** distribution(Intuition: Number of successes are X_i before the first failure occurs(and then everything resets!))
$X_i \sim NB(1,p)$ which is basically a geometric distribution. And hence $P(X_i = k) = (1-p)^k p$

Problem 3: (c)

From Theorem 11.18, if $X_1, X_2 \dots X_n$ are i.i.d then

$$\lim_{n\to\infty} P(Y_n < a_n + b_n y) = e^{-u(y)} \tag{3c.1}$$

where,

$$\lim_{n\to\infty} n\{1 - F(a_n + b_n y)\} = u(y) \tag{3c.2}$$

We have: $F(X) = P(X \le x) = 1 - p^x$
Now ,

$$\begin{aligned} \lim_{n\to\infty} P\{b_n(Y_n - a_n) \le z\} &= \exp(-\exp(z)) \\ \lim_{n\to\infty} P\{(Y_n \le \frac{1}{b_n}z + a_n\} &= \exp(-\exp(z)) \end{aligned} \tag{3c.3}$$

Thus, comparing coefficients in $(3c.1), (3c.2), (3c.3)$, we have:

$$\begin{aligned} \lim_{n\to\infty} n(1 - F(a_n + \frac{1}{b_n}z)) &= \exp(-z) \\ \lim_{n\to\infty} n(1 - (1 - p^{a_n + \frac{1}{b_n}z})) &= \exp(-z) \\ \lim_{n\to\infty} np^{a_n + \frac{1}{b_n}z} = \exp(-z) \lim_{n\to\infty} (np^{a_n})(p^{\frac{1}{b_n}}e)^z & = 1 \end{aligned}$$

Satisfying which requires:

$$\begin{aligned} np^{a_n} &= 1 \\ 1/p^{a_n} &= n \\ \implies a_n &= \log_{1/p}(n) \end{aligned}$$

And,

$$\begin{aligned} (p^{\frac{1}{b_n}}e)^z &= 1 \\ z(\log(p^{\frac{1}{b_n}}) + 1) &= 0 \\ \implies b_n &= \log(1/p) \end{aligned}$$

Thus,

$$\begin{aligned} a_n &= \log_{1/p}(n) \\ b_n &= \log(1/p) \end{aligned}$$

Problem 3: (d)

$E(M_n) = n(1-p) = nq$ Now, with similar calculations as in the last part it is possible to show that this follows an extreme value distribution:

$$P\{\log(1/p)(Y_n - \log_{1/p}(nq)) < z\} \to \exp(-\exp(-z))$$

Problem 3: (e)

Part(i)

$$\begin{aligned}
E(s(A,B)) &= \eta_0\gamma_0 \times s(0,0) + \eta_1\gamma_0 \times s(1,0) + \eta_0\gamma_1 \times s(0,1) + \eta_1\gamma_1 \times s(1,1) \\
&= \frac{1}{3} + \frac{1}{6} - 2\frac{1}{6} - 2\frac{1}{3} \\
&= -\frac{1}{2} && < 0
\end{aligned}$$

Part (ii)
To calculate such a λ such that $\lambda(R_n - \ln(K_n))$ has extreme value distribution, we find roots of $E(e^{xS(A,B)}) = 1$
$p = \sum_a \eta_a\gamma_a = \frac{1}{2}$
So, $pe^{\lambda} + (1-p)e^{-2\lambda} = 1 \implies e^{\lambda} + e^{-2\lambda} = 2$
thus, $t^2 + t - 2 = 0$ $t = \frac{-1\pm\sqrt{5}}{2}$
And since $\lambda > 0 \implies \lambda = \exp(\frac{1+\sqrt{5}}{2})$
Part (iii)
In aligned part: $p_{ab} = \eta_a\gamma_b e^{\lambda s(a,b)}$

$$\begin{aligned}
p_{00} &= \frac{1}{3}\frac{1+\sqrt{5}}{2} \\
p_{01} &= \frac{1}{3}(\frac{1+\sqrt{5}}{2})^{-2} \\
p_{10} &= \frac{1}{6}(\frac{1+\sqrt{5}}{2})^{-2} \\
p_{11} &= \frac{1}{6}(\frac{1+\sqrt{5}}{2})
\end{aligned}$$

Problem 4

Problem 4: (a)

$$P(n \text{ individuals with allele A}) = \binom{N}{n} f_A^n (1-f_A)^{N-n}$$

Problem 4 continued on next page. . .

Problem 4: (b)

Coverage is λ, thus the probability that this particular locus does NOT get sequenced $= \exp(-c)$

$$\begin{aligned}
P(\text{At least one read with allele 'A' AND 'a'}) &= 1 - P(\text{ Zero reads with allele 'A' OR 'a'}) \\
&= 1 - P(\text{Zero reads with 'a'}) \\
&- P(\text{Zero reads with allele 'A'}) \\
&+ P(\text{Zero reads with allele 'A' and 'a'}) \\
&= 1 - (\exp(-\lambda))^n - (\exp(-\lambda))^{N-n} + (\exp(-\lambda))^n \\
&= 1 - e^{-n\lambda} - e^{-\lambda(N-n)} + e^{-\lambda N}
\end{aligned}$$

Problem 4: (c)

For the locus to be declared polymorphic, there should be *at least* one read with 'A' and *at least* one read with allele 'a'.

$$\begin{aligned}
P(\text{locus is polymorhic}) &= 1 - P(\text{locus is not polymorphic}) \\
&= 1 - P(\text{ Zero reads with allele 'A' AND 'a'}) \\
&= 1 - e^{-n\lambda} - e^{-\lambda(N-n)} + e^{-\lambda N}
\end{aligned}$$

Problem 5

Problem 5: (a)

$$L(n_{ii',jj'}|p_{00}p_{01}p_{10}p_{11}) = \prod_{ii',jj'} \left(\alpha_{ii'jj'}p_{ij}p_{ij'}\right)^{n_{ii'jj'}}$$

$$\alpha_{ii'jj'} = \begin{cases} 1 & (i = i'; j = j') \\ 2 & (i = i'; j \neq j' \; OR \; i \neq i'; j = j') \\ 4 & (i \neq i'; j \neq j') \end{cases}$$

and $i, i', j, j' \in \{0, 1\}$

Problem 5: (b)

Missing Data?
Let's look at the haplotypes:

	00	01	11
00	(00,00)	(00,01)	(01,01)
01	(00,10)	(00,11);(01,10)	(00,10)
11	(10,10)	(10,11)	(11,11)

Ambiguity in haplotypes occur whenever any of loci 'A,B' is heterozygous or both are heterozygous. $n_{10,10}$ in this case gives rise to twp haplotype pairs: $(11, 00); (10, 01)$ and We cannot directly determine the exact count from the genotype information. In otherwords the haplotype counts $n_{(11/00)}$ and $n_{(10/01)}$ are the missing data.
Thus, missing data: $n_{00/11}$ and $n_{01/01}$.

We assume there N individuals and hence there are $2N$ haplotypes.

Observed data: $Y = (n_{0000}, n_{1100}, n_{1111}, n_{0001}, n_{1101}, n_{0111}, n_{0101})$

Missing Data: $n_{00/11}$ and $n_{01/10}$

We construct complete data as the haplotype counts:
Complete Data: $n_{00}, n_{01}, n_{10}, n_{11}$

Parameters: $\theta = (p_{00}, p_{01}, p_{10}, p_{11})$

and hence the Complete data likelihood is given by:

$$g(n_{00}, n_{01}, n_{10}, n_{11}|\theta) = \frac{2N}{n_{00}!n_{01}!n_{10}!n_{11}!} p_{00}^{n_{00}} p_{01}^{n_{01}} p_{10}^{n_{10}} p_{11}^{n_{11}}$$

Problem 5: (c)

In the E step. we perform (m^{th} step):

$$\hat{n_{00}} = E[n_{00}|Y,\theta_m]$$
$$\hat{n_{01}} = E[n_{01}|Y,\theta_m]$$
$$\hat{n_{10}} = E[n_{10}|Y,\theta_m]$$
$$\hat{n_{11}} = E[n_{11}|Y,\theta_m]$$

where $\theta_m = (p_{00}^{(m)}, p_{01}^{(m)}, p_{10}^{(m)}, p_{11}^{(m)})$
Just consider n_{00} for now.

$$\begin{aligned} n_{00} &= E[n_{00}|Y,\theta_m] \\ &= 2n_{0000} + n_{0010} + n_{0100} + E[n_{00/11}|Y,\theta_m] \end{aligned}$$

where the last term comes because the 00 haplotype can also come from the ambiguos we highlighted in the table above $(11,00);(10,01)$
Now, we need to consider:

$$\begin{aligned} E[n_{00/11}|Y,\theta_m] &= n_{0101} P(00/11|01/10, 00/11) \\ &= n_{0101} \times (\frac{2p_{00}p_{11}}{2p_{00}p_{11} + 2p_{01}p_{10}}) \end{aligned}$$

Where the latter term comes out from the conditional probability of observing 01/10 haploltype given it is coming from a heterozygous subpopulation at both A,B
Thus, the E step gives us:

$$\hat{n_{00}} = 2n_{0000} + n_{0010} + n_{0100} + n_{0101}\frac{p_{00}p_{11}}{p_{00}p_{11} + p_{01}p10}$$
$$\hat{n_{01}} = 2n_{0011} + n_{0001} + n_{0100} + n_{0101}\frac{p_{01}p_{10}}{p_{00}p_{11} + p_{01}p10}$$
$$\hat{n_{10}} = 2n_{1100} + n_{1000} + n_{1101} + n_{0101}\frac{p_{10}p_{01}}{p_{00}p_{11} + p_{01}p10}$$
$$\hat{n_{11}} = 2n_{1111} + n_{1010} + n_{1110} + n_{0101}\frac{p_{11}p_{00}}{p_{00}p_{11} + p_{01}p10}$$

Problem 5: (d)

$$g(n_{00}, n_{01}, n_{10}, n_{11}|\theta) = \frac{2N}{n_{00}!n_{01}!n_{10}!n_{11}!} p_{00}^{n_{00}} p_{01}^{n_{01}} p_{10}^{n_{10}} p_{11}^{n_{11}}$$

At the M step, we maximise the likelihood function (g) with respect to θ_m, since it is a multinomial, and we know the the MLE for a multinomial is simply given by the ratio of $p_x = \frac{n_x}{N}$
we get:

$$p_{00} = \frac{\hat{n_{00}}}{2N}$$
$$p_{01} = \frac{\hat{n_{01}}}{2N}$$
$$p_{10} = \frac{\hat{n_{10}}}{2N}$$
$$p_{11} = \frac{\hat{n_{11}}}{2N}$$

Problem 5: (e)

See Appendix 2.

Problem 5: (f)

Challenge: If there are L loci, there are 2^L haplotypes and hence the EM algoriothm steps will grow exponentially.
Approach: We can take a Monte Carlo approach, sampling few n loci out of L in the beginning, estimate their frequency till convergence using and then use this data to further estimate the rest $L-n$ frequencies.