20W-COMSCICM124-1 MidtermVersion2

TOTAL POINTS

8.0007 / 10

QUESTION 1

1 1 **1 / 1**

✓ **+ 1 pts Correct**

+ 0 pts not try

+ 0.02 pts wrong

QUESTION 2

2 2 **1 / 1**

✓ **+ 1 pts Correct**

+ 0 pts Click here to replace this description.

+ 0.02 pts Click here to replace this description.

QUESTION 3

3 3 **1 / 1**

✓ **+ 1 pts Correct**

+ 0 pts no try

+ 0.02 pts wrong

QUESTION 4

4 4 **1 / 1**

✓ **+ 1 pts Correct**

+ 0 pts no try

+ 0.02 pts wrong

QUESTION 5

5 5 **1 / 1**

✓ **+ 1 pts Correct**

+ 0 pts no try

+ 0.02 pts wrong

QUESTION 6

6 6 **0.0007 / 1**

+ 1 pts Correct

+ 0 pts no try

+ 0.02 pts wrong

✓ **+ 0.0007 pts partial 70**

+ 0.0003 pts partial 30

QUESTION 7

7 7 **1 / 1**

✓ **+ 1 pts Correct**

+ 0 pts no try

+ 0.02 pts wrong

QUESTION 8

8 8 **1 / 1**

✓ **+ 1 pts Correct**

+ 0 pts no try

+ 0.02 pts wrong

QUESTION 9

9 9 **1 / 1**

✓ **+ 1 pts Correct**

+ 0 pts no try

+ 0.02 pts wrong

+ 0.0007 pts partial 70

+ 0.0003 pts partial 30

QUESTION 10

10 10 **0 / 1**

+ 0 pts Correct

+ 0 pts no try

✓ **+ 0 pts wrong**

CM124 CM224: Machine Learning Applications in Genetics Winter 2020

Midterm Exam, Version 2

- **Each question must have an explanation in the box below the answer selection.** This explanation must convince the reader that you had a valid reason for your answer, and that your choice was not random. If you do not justify your answer, the problem will be marked as incorrect.
- The exam is open notes, but electronics of any kind are not allowed except for simple calculators that cannot access the internet.
- **There are 10 questions in total.** The last question is extra credit. If you correctly answer this question, you get 10 extra points. If you incorrectly answer this question, you will get 5 points deducted from your total points.
- **Each wrong answer reduces the total point by 2 points.**
  - If you correctly answer 9 of the 9 questions your grade will be 100.
  - If you correctly answer 8 of the 9 questions your grade will be 94.
  - If you correctly answer 7 of the 9 questions your grade will be 87.
  - If you correctly answer 6 of the 9 questions your grade will be 79.
  - If you correctly answer 5 of the 9 questions your grade will be 70.
  - If you correctly answer 4 of the 9 questions your grade will be 60.
  - If you correctly answer 3 of the 9 questions your grade will be 45.
  - If you correctly answer 2 of the 9 questions your grade will be 30.
  - If you correctly answer 1 of the 9 questions your grade will be 15.
  - If you correctly answer 0 of the 9 questions your grade will be 0.

Name and ID:

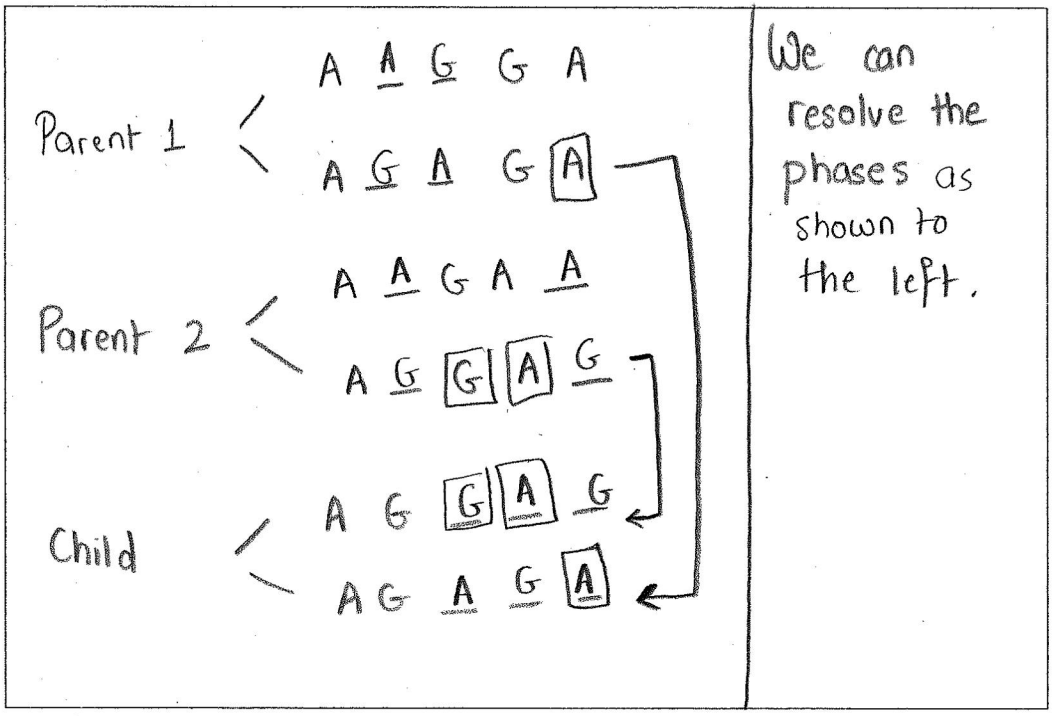
1. **Trio phasing.** Suppose you have the following trio with the genotypes as below (where H denotes a heterozygous base)

$$\begin{aligned} \text{Parent1} &= \text{AHHGA} \\ \text{Parent2} &= \text{AHGAH} \\ \text{Child} &= \text{AGHHH} \end{aligned}$$

Assume there is no recombination and no mutations, and that all bases are either A or G. What is the phase of the child in this trio?

- ■ AGAGA, AGGAG
- ☐ AGAAA, AGGGG
- ☐ The phase of the child cannot be resolved.
- ☐ The phase of the child can be resolved, but is not A or B.
- ☐ None of the above.

Briefly explain your answer in the box.



2. **Clark's Algorithm.** You have sequenced the genotypes HHAG, HHAH, and AGHH (in that order) for a certain region of the genome in a population for which the haplotypes AGAG, AGGG, and AGAA are known to exist in this population. Use Clark's algorithm (as described in class) to phase these genotypes. Which of the following haplotypes is the last one added to your known set?

☐ AGGA
☐ GAAA
■ GAAG
☐ There are no haplotypes added to the known set.
☐ None of the above.

Briefly explain your answer in the box.

HHAG: AGAG ← known; GAAG ← added
HHAH: GAAG ← already added; AGAA ← known
AGHH: AGAA ← known; AGGG ← known

So we add GAAG

3. **Haplotype frequency estimation with Lagrangian.** Suppose that we observe counts of 5, 10, and 10 for three haplotypes in a population with frequencies p_1, p_2, and 0.5, with the latter being known due to extensive studies of this population. Assuming a multinomial likelihood function for the frequency parameters based on the observed counts, what is the value of p_1 given by its maximum likelihood estimator? Write your answer as a fraction in the box below.

$p_1 = \frac{1}{6}$

Briefly explain your answer in the box.

$c_1 = 5 \quad p_1 = ?$

$c_2 = 10 \quad p_2 = ?$

$c_3 = 10 \quad p_3 = 0.5$

Likelihood can be maximized as argmax $p_1^{c_1} p_2^{c_2} p_3^{c_3}$

Maximize $p_1^{c_1} p_2^{c_2}$ since p_3 is fixed.

$$c_1 \log p_1 + c_2 \log p_2 + \lambda(p_1 + p_2 - 0.5)$$

$\frac{\partial F}{\partial p_1} = \frac{c_1}{p_1} + \lambda = 0 \quad \frac{\partial F}{\partial p_2} = \frac{c_2}{p_2} + \lambda = 0$

$\frac{c_1}{p_1} = \frac{c_2}{p_2} \quad p_2 = \frac{c_2}{c_1} p_1 = 2p_1$

$p_1 + p_2 = 0.5$

$3p_1 = 0.5$

$p_1 = \frac{0.5}{3} = \frac{1}{6}$

$p_2 = \frac{2}{6}$

4. **Gradient Ascent.** Suppose that we want to find the maximum likelihood estimators of haplotype frequencies p_1 and p_2 given their observed counts of 5 and 10, respectively, using gradient ascent as described in class. Use the log-likelihood function $\sum_i c_i log(p_i)$ and the constraint $p_1 + p_2 = 1.0$. Assume that there is no missing or masked data and that our initial guess is $p_1 = p_2 = 0.5$ and our step size is $\epsilon = 0.01$. After one iteration of gradient ascent, what are the new values for p_1 and p_2? Write the two numbers as decimals with 2 digits (for example, 0.01) in the box.

$P_1 = 0.45$
$P_2 = 0.55$

Briefly explain your answer in the box.

$$g(p) = \sum_i c_i \log(p_i)$$

$$\frac{\partial g}{\partial p_i} = \frac{c_i}{p_i} \qquad \lambda = \frac{\sum_{i=1}^{m} \frac{\partial g}{\partial p_i}}{m} = \frac{\frac{5}{P_1} + \frac{10}{P_2}}{2} = \frac{5}{2}\left(\frac{P_2 + 2P_1}{P_1 P_2}\right)$$

$$\frac{\partial g}{\partial p_1} = \frac{5}{P_1} \qquad \frac{\partial g}{\partial p_2} = \frac{10}{P_2} \qquad = \frac{5}{2}\left(\frac{0.5+1}{0.5 \times 0.5}\right)$$

$$p = (0.5, 0.5)$$

$$u = \left(\frac{\partial g}{\partial p_1} - \lambda, \frac{\partial g}{\partial p_2} - \lambda\right) = (10-15, 20-15) = (-5, 5)$$

$$p' = p + \varepsilon u = (0.5, 0.5) + 0.01(-5, 5)$$

$$= (0.5 - 0.05, 0.5 + 0.05)$$

$$= (0.45, 0.55)$$

5. **Frequencies of masked haplotypes.** Assume that haplotypes 100, 110, and 111 have frequencies 0.5, 0.25, and 0.25 in the population, respectively. Suppose that we sample two haplotypes from these possible haplotypes, but then some bases are independently masked with probability 0.5, such that we observe a missing value denoted as * instead of the actual base. What is the likelihood of observing haplotypes 100 and 1**? Write your estimation as a fraction in the box (you do not need to simplify).

$\frac{1}{2^7} = \frac{1}{128}$ $= 0.0078125$

Briefly explain your answer in the box.

100 - 0.5
110 - 0.25
111 = 0.25

The other one could be any. Since we need to mask the last two positions, the probability of 1** is

$1 \times (1-0.5) \times 0.5 \times 0.5 = 0.125$
(not mask, mask, mask)

Probability of 100 is $0.5 \times (1-0.5)^3 = 0.0625$
(0.5: picking 100; $(1-0.5)^3$: not mask)

Since both are independent, the joint probability is

$0.125 \times 0.0625 = 0.0078125 = (0.5)^7$

$= \frac{1}{2^7}$

6. **Recombination rate for haplotypes.** Suppose you have one person with two haplotypes, 00 and 11, and a probability r for a recombination between the 2 SNPs. What is the probability that this person transmits the haplotype 01 to their child? Write your answer in the box.

r

Briefly explain your answer in the box.

r
0|0
1|1

When they recombine we get 01, 01. So both haplotypes after recombination give the desired haplotype 01.
The probability from recombination is r, so the solution is r.

7. **EM algorithm.** Suppose there are three genotypes 220, 101, and 002. To compute the haplotype frequencies for this genotype data, you need to apply the EM algorithm. In the first iteration of the EM algorithm, assume your initial guess for the haplotype frequencies are $p_{hi} = 1/8$ (there are 8 possible haplotypes). Use the objective function Q in the EM algorithm as defined in class. Now, to solve for the haplotype frequencies after one iteration of EM, you will need to compute the derivative of Q with respect to p_{hi}. Write the derivative $\partial Q/\partial p_{001}$ for the haplotype 001 in the box.

$$\frac{\partial Q}{\partial p_{001}} = \frac{2a^{002}_{001,001}}{p_{001}} + \frac{a^{101}_{001,100}}{p_{001}}$$

Briefly explain your answer in the box.

220 — 110 / 110

101 — 101 / 000 / 100 / 001

002 — 001 / 001

$$Q(\vec{p}|p^t) = \sum_g \sum_{(h_1,h_2)\in(g)} a^g_{h_1,h_2} \log(p_{h_1} p_{h_2})$$

$$\frac{\partial Q}{\partial p_{001}} = \frac{\partial}{\partial p_{001}} \left(a^{002}_{001,001} \log(p^2_{001}) + a^{101}_{001,100} \log(p_{001} p_{100}) \right)$$

$$= a^{002}_{001,001} \frac{2p_{001}}{p^2_{001}} + a^{101}_{001,100} \frac{p_{100}}{p_{001} p_{100}}$$

$$= \frac{2a^{002}_{001,001}}{p_{001}} + \frac{a^{101}_{001,100}}{p_{001}}$$

$$= \frac{2}{\frac{1}{8}} + \frac{\frac{1}{2}}{\frac{1}{8}} = 16 + 4 = 20$$

$a^{002}_{001,100} = 1$

$a^{101}_{100,001} = \frac{1}{2}$

8. **Missing data in EM algorithm.** Genotyping is imperfect, so usually there is some missing data. Assume that missing data occurs at random and is independent of the haplotype. Suppose there are three people with genotypes 22*, 101, and 002, where * denotes a missing base. Assume that your initial guess for the haplotype frequencies are $p_{hi} = 1/8$. Now do one iteration of the EM algorithm to estimate the frequency of haplotype 111. Write your answer as a fraction in the box.

$\frac{1}{6}$

Briefly explain your answer in the box.

22 * → 110 $\frac{1}{3}$, 110 $\frac{1}{3}$, 111 $\frac{1}{3}$, 111 $\frac{1}{3}$

101 → 100 $\frac{1}{2}$, 001 $\frac{1}{2}$; 101, 000

002 → 001, 001 (1)

$P_{110} = \frac{2 \times \frac{1}{3} + \frac{1}{3}}{6} = \frac{1}{6}$

$P_{111} = \frac{2 \times \frac{1}{3} + \frac{1}{3}}{6} = \frac{1}{6}$

$P_{100} = \frac{\frac{1}{2}}{6} = \frac{1}{12}$

$P_{001} = \frac{\frac{1}{2} + 1 \times 2}{6} = \frac{5}{12}$

$P_{000} = \frac{1}{12}$

$P_{101} = \frac{1}{12}$

9. **Maximum likelihood estimation.** Suppose you observe a series of independent and identically distributed (IID) random variables $X_1, .., X_n \sim \text{Erlang}(k, \lambda)$. You do not need to know what the Erlang distribution is. Its PDF is defined as follows:

$$f(x; k, \lambda) = \frac{\lambda^k x^{k-1} e^{-\lambda x}}{(k-1)!}$$

which is defined for $x \in [0, \infty)$, with the parameter ranges $k \in \{1, 2, 3, 4, ...\}$ and $\lambda \in (0, \infty)$. What is the maximum likelihood estimate for λ? Write your answer in the box below.

$$\lambda = \frac{kn}{\sum_{i=1}^{n} x_i}$$

Briefly explain your answer in the box.

$$\log \prod_{i=1}^{n} \frac{\lambda^k x_i^{k-1} e^{-\lambda x_i}}{(k-1)!} = \sum_{i=1}^{n} \left(k \log \lambda + (k-1) \log x_i - \lambda x_i - \log((k-1)!) \right)$$

$$\frac{\partial \ell}{\partial \lambda} = \sum_{i=1}^{n} \left(\frac{k}{\lambda} - x_i \right) = 0$$

$$= \sum_{i=1}^{n} \frac{k}{\lambda} - \sum_{i=1}^{n} x_i = \frac{kn}{\lambda} - \sum_{i=1}^{n} x_i$$

$$\lambda = \frac{kn}{\sum_{i=1}^{n} x_i}$$

10. **Extra credit question.** If you correctly answer this question, you get 10 extra points. If you incorrectly answer this question, 5 points will be deducted from your total points.

Assume that there is no recombination and no missing data. Assume that we have a mother, father and child trio, and that their genotypes are mother: 102, father: 110, and child: 111. In the box below, write the EM objective function to estimate the haplotype frequencies in the population based on this sample. You do not need to solve this objective function.

$$Q(\vec{p}|\vec{p}^t) = a^{102}_{001,101}\log(P_{001}P_{101}) + a^{110}_{110,000}\log(P_{110}P_{000}) + a^{110}_{100,010}\log(P_{100}P_{010}) + a^{111}_{110,001}\log(P_{110}P_{001}) + a^{111}_{101,010}\log(P_{101}P_{010})$$

Briefly explain your answer in the box.

mother : 102 — 001, 101

father : 110 — 110, 000, 100, 010

child : 111 — ~~111~~ ~~000~~ no(111), 110, 001, 101, 010, ~~011~~ ~~100~~ no(011)

$$L(P_1 \cdots P_n; G) = (P_{001}P_{101})(P_{110}P_{000} + P_{100}P_{010}) \times (P_{110}P_{001} + P_{101}P_{010})$$

$$Q(\vec{p}|p^t) = a^{102}_{001,101}\log(P_{001}P_{101}) + a^{110}_{110,000}\log(P_{110}P_{000}) + a^{110}_{100,010}\log(P_{100}P_{010}) + a^{111}_{110,001}\log(P_{110}P_{001}) + a^{111}_{101,010}\log(P_{101}P_{010})$$