CM124 - Spring 2018 - Midterm

April 25, 2018

Name: [redacted]

Student ID: [redacted]

- Exam questions are multiple choice, only one answer is correct.
- Any questions without an answer selection, or multiple answer selections, will be marked as incorrect.
- Each question must have a written justification in the box below the answer selection. This justification must convince the reader that you had a valid reason for choosing the answer you chose, and that your choice was not random. If you do not justify your answer, or if your justification is not convincing, the problem will be marked as incorrect.
- If you answer all questions correctly your grade will be 100.
- If you answer correctly 5 out of 6 questions your grade will be 95.
- If you answer correctly 4 out of 6 questions your grade will be 90.
- If you answer correctly 3 out of 6 questions your grade will be 85.
- If you answer correctly 2 out of 6 questions your grade will be 75.
- If you answer correctly 1 out of 6 questions your grade will be 60.
- If you answer correctly 0 out of 6 questions your grade will be 0.
- The exam is open notes, but electronics of any kind are not allowed.

1. **MLE of Allele Frequency.** Suppose you have the following alleles and their respective counts:
0 - 600
1 - 200

What is the maximum likelihood estimate of the frequency of the 1 allele?

Select from the following options and justify briefly in the box why you choose your answer.

(a) 0.1
(b) 0.75
(c) 0.9
(d) 0.25

The MLE in class shows that it's $\frac{m}{n}$ where m = # 1s, n = total # alleles.

So: $\frac{m}{n} = \frac{200}{600+200} = \frac{200}{800} = \frac{2}{8} = \frac{1}{4} = 0.25$ (D)

2. **MLE question of Haplotype Frequency.**

Suppose you have the following haplotypes and their respective counts:

1001 - 300
1101 - 1200
0011 - 400
0101 - 100

What is the maximum likelihood estimate of the haplotype 1101?

Select from the following options and justify briefly in the box why you choose your answer.

(a) 0.1
(b) 0.25
(c) 0.6
(d) 0.2

similar to previous problem, the MLE is $\frac{m}{n}$, as derived in class, where m is count of 1101 haplotype & n = total counts of all haplotypes. so:

$\frac{m}{n} = \frac{1200}{1200+300+400+100} = \frac{1200}{2000} = \frac{12}{20} = \frac{6}{10} = 0.6$

1500 500

C

3. **Trio Phasing.** Suppose you have the following trio with the genotypes as below

$$\begin{aligned} \text{Parent1} &= 122102 \\ \text{Parent2} &= 011120 \\ \text{Child} &= 112011 \end{aligned}$$

Assume there is no recombination. What is the phase of the child in this trio?

Select from the following options and justify briefly in the box why you choose your answer.

(a) 011001, 101010
(b) 111001, 001010
(c) 011001, 111010
~~(d) 110010, 001010~~

we discern:

Parent1: 122102 < ?11?01] ?11?01] To Child

Parent2: 011120 < 0???10] To child
0???10

child: 112011 < ??10??] From Parent 1
??10??] From Parent 2

we discern:

Parent1: 1 1 1 0 0 1 *
0 1 1 1 0 1

Parent2: 0 0 1 0 1 0 *
0 1 0 1 1 0

child 1 1 1 0 0 1 P1
0 0 1 0 1 0 P2

B

4. **Clark's algorithm.**
Suppose you have the following genotypes:

AGAAA
AGHAA AGGAA
GAGHA
HHGHA
AHAAA
HHHHA

Perform Clark's algorithm, starting from the first genotype. what is the final haplotype(s) that were added to your known set.

Select from the following options and justify briefly in the box why you choose your answer.

(a) AAAAA
(b) GGGGA
(c) AGAAA
(d) AGGAA

A is 0 > (Homozygous)
G is 2
H is 1 (Heterozygous)

knowns!
AGGAA
AGAAA
GAGAA
GAGGA
AAAAA

AGHAA < AGGAA / AGAAA

GAGHA < GAGAA / GAGGA

HHGHA < ??G?A / ??G?A → can be solved with: AGGAA / GAGGA → HHGHA

AHAAA < AAAAA / AGAAA ★ already added

HHHHA < ????A / ????A → Can be solved with: AGAAA / GAGGA → HHHHA

Since we only use knowns to find the unknowns, the last one we added to the known bank was AAAAA

A

$1 \cdot 4 + \frac{1}{2} \cdot 4 = 4 + 2 = 6$

5. **Haplotype phasing with the EM algorithm.**

Suppose you have genotypes $G = \{01210, 10222, 00110\}$. Perform one round of the EM algorithm for haplotype phasing (assume that the haplotypes are equally probable $p_1 = p_2 ... = p_n = 1/n$). After one round of EM, what haplotype(s) have the highest estimated probability?

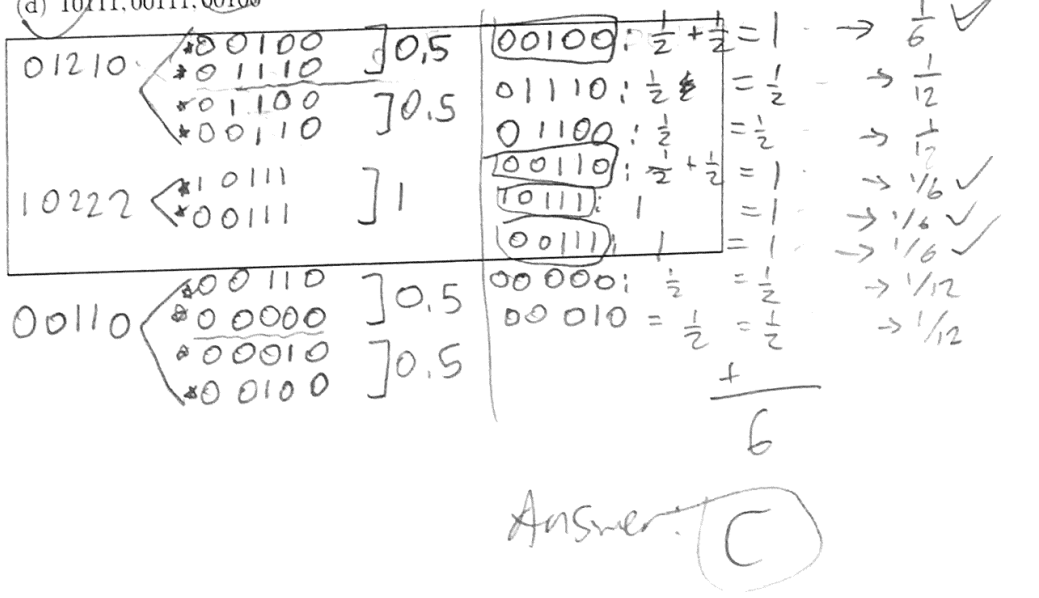
Select from the following options and justify briefly in the box why you choose your answer.

(a) 00110, 10111

(b) 00100, 00111, 00010, 00000

(c) 10111, 00111, 00110, 00100

(d) 10111, 00111, 00100



Answer: (C)

Actual answer

$p' = p + \epsilon u \qquad u - \tilde{\nabla} f = \left[\frac{dL}{dp_1} - \frac{\sum_{i=1}^{n} \frac{dL}{dp_i}}{n} \right.$

$\log_x L(p_1, \ldots, p_n) = \sum_{i=1}^{n} c_i \log(p_i)$

$p_1 = p_2 = \ldots = \frac{1}{n}$

$p_i + \epsilon \frac{c_i}{p_i} \qquad p_j + \epsilon \frac{c_j}{p_j} \qquad p_i$

$p_i - p_j =$

6. Likelihood optimization

$\frac{dL}{dp_1} = \frac{c_1}{p_1} = \frac{c_1}{1/n} = c_1 n$

$\sum_{i=1}^{n} c_i n = \sum_{i=1}^{n} c_i = C$

$p_i^{(1)} - p_j^{(1)} = \frac{1}{n} - \frac{1}{n} + \epsilon(c_i n - C) - \epsilon(c_j n - C) = \epsilon(c_i n - c_j n) = \epsilon(c_i - c_j)$

Consider the likelihood function of the haplotype frequency estimation problem with no missing data. In this setting, the parameters are the haplotype allele frequencies $p_1, \ldots, p_n$, satisfying $\sum_{i=1}^{n} p_i = 1$, and $p_i \geq 0$, and the data is the haplotype counts $c_1, \ldots, c_n$, and let $c = c_1 + \ldots + c_n$. We showed in class that the likelihood function $L(p_1, \ldots, p_n)$ satisfies

$$\log L(p_1, \ldots, p_n) = \sum_{i=1}^{n} c_i \log(p_i)$$

Assume we are running a gradient ascent with projections algorithm, and we start from the guess $p_1^{(0)} = p_2^{(0)} = \ldots = p_n^{(0)} = \frac{1}{n}$. Let $p_1^{(1)}, \ldots, p_n^{(1)}$ be the next point that the algorithm reaches. Which of the following is true:

(a) There is $\epsilon > 0$ such that for every i, j we have $p_i^{(1)} - p_j^{(1)} = (c_i - c_j)\epsilon$.

(b) There is $\epsilon > 0$ such that for every i, j we have $p_i^{(1)} - p_j^{(1)} = (c_j - c_i)\epsilon$.

(c) There is $\epsilon > 0$ such that for every i we have $p_i^{(1)} = (c_i - \frac{c}{n})\epsilon$.

(d) There is $\epsilon > 0$ such that for every i, j we have $p_i^{(1)} + p_j^{(1)} = \frac{2}{n} + (c_i + c_j)\epsilon$.

Let current point $= Cp$

Select from the above options and justify briefly in the box why you choose your answer.

Let $g(p) = \sum_i^n c_i \log(p_i)$, s.t. $\sum p_i = 1, p_i \geq 0$

Cp: $p = Cp_1, \ldots, p_n)$

$u = \left(\frac{dg}{dp_1} - \lambda, \ldots, \frac{dg}{dp_n} - \lambda \right) \qquad \frac{dg}{dp_i} = \sum_i^n \frac{c_i}{p_i}$

For small $\epsilon > 0$:

$p' = p + \epsilon u = (p_1, \ldots, p_n) + \epsilon\left(\frac{c_1}{p_1} - \lambda, \frac{c_2}{p_2} - \lambda, \ldots, \frac{c_n}{p_n} - \lambda \right)$

where $\lambda = \frac{\sum_i^n \frac{dg}{dp_i}}{n}$

$= \left(p_1 + \epsilon\left(\frac{c_1}{p_1} - \lambda\right), \ldots, p_n + \epsilon\left(\frac{c_n}{p_n} - \lambda\right) \right)$

$p' = \left(p_1 + \epsilon\left(\frac{c_1}{p_1} - \frac{\sum \frac{dg}{dp_i}}{n}\right), \ldots, p_n + \epsilon\left(\frac{c_n}{p_n} - \frac{\sum_i^n \frac{dg}{dp_i}}{n}\right) \right) = \left(p_1^{(1)}, \ldots, p_i^{(1)}, \ldots, p_j^{(1)}, \ldots, p_n^{(1)} \right)$

For each i, j: $p_i^{(1)} - p_j^{(1)} = p_i^{(0)} + \epsilon\left(p_i^{(0)} + \epsilon \frac{c_i}{p_i^{(0)}} - \frac{\sum \frac{dg}{dp_i}}{n} \right) -$

$\left[p_j^{(0)} + \epsilon\left(p_j^{(0)} + \epsilon \frac{c_j}{p_j^{(0)}} - \frac{\sum \frac{dg}{dp_i}}{n} \right) \right]$

$\approx \epsilon(c_i) - \epsilon(c_j) = \epsilon(c_i - c_j)$