b

CM124 - Spring 2018 - Midterm

April 25, 2018

[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

$$\hat{p} = \frac{c_1}{n} = \frac{200}{600+200} = \frac{1}{4} = 0.25$$

where
$\hat{p}$ = MLE for allele frequency, as shown in class
c_1 = count of "1" allele
n = total of allele counts

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

$$\hat{p} = \frac{c_{1101}}{n} = \frac{1200}{300+1200+400+100} = \frac{3}{5} = 0.6$$

Where
$\hat{p}$ = MLE for haplotype frequency, as shown in class
c_{1101} = count of "1101" haplotype
n = total of haplotype counts

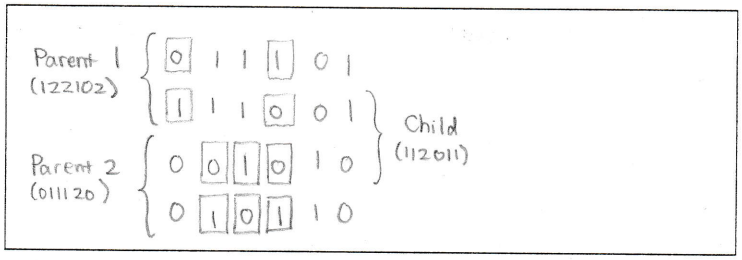
3. **Trio Phasing.** Suppose you have the following trio with the genotypes as below

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

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



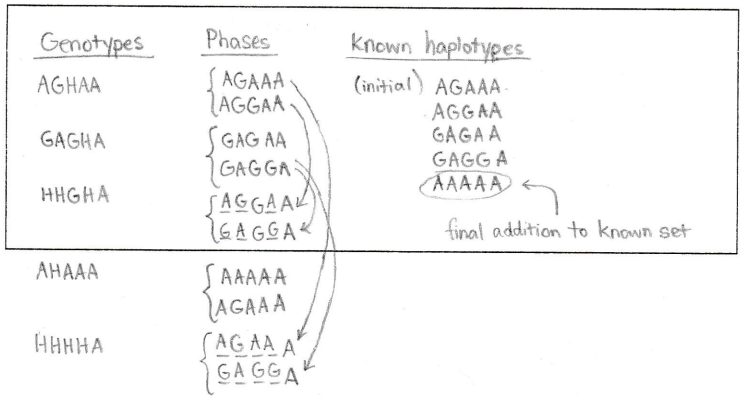
4. **Clark's algorithm.**
Suppose you have the following genotypes:

AGHAA
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



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

Genotypes	Phases	Probabilities
01210	{00100 △, 01110	$\frac{(\frac{1}{8})^2}{2(\frac{1}{8})^2} = \frac{1}{2}$
	{01100, 00110 *	" $= \frac{1}{2}$
10222	{00111, 10111	$\frac{(\frac{1}{8})^2}{(\frac{1}{8})^2} = 1$
00110	{00000, 00110 *	$\frac{(\frac{1}{8})^2}{2(\frac{1}{8})^2} = \frac{1}{2}$
	{00100 △, 00010	" $= \frac{1}{2}$

Haplotypes

△ 00100 $\frac{1}{8} \to k(\frac{1}{2}+\frac{1}{2}) = \frac{1}{6}$

01110 $\frac{1}{8} \to k(\frac{1}{2}) = \frac{1}{12}$

01100 $\frac{1}{8} \to k(\frac{1}{2}) = \frac{1}{12}$

* 00110 $\frac{1}{8} \to k(\frac{1}{2}+\frac{1}{2}) = \frac{1}{6}$

00111 $\frac{1}{8} \to k(1) = \frac{1}{6}$

10111 $\frac{1}{8} \to k(1) = \frac{1}{6}$

00000 $\frac{1}{8} \to k(\frac{1}{2}) = \frac{1}{12}$

00010 $\frac{1}{8} \to k(\frac{1}{2}) = \frac{1}{12}$

$k := \frac{1}{2|G|} = \frac{1}{6}$

6. **Likelihood optimization**

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

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

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 $p_i^{(1)} + p_j^{(1)} = \frac{2}{n} + (c_i + c_j)\epsilon$.

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

$p' = p + \epsilon u$

$p_i^{(1)} - p_j^{(1)} = [p_i^{(0)} + \epsilon u_i] - [p_j^{(0)} + \epsilon u_j]$

$p_i^{(1)} - p_j^{(1)} = (u_i - u_j)\epsilon$

$u_i, u_j \sim c_i, c_j$