CSCI 2820 HW3: Haplotype Phasing and GWAS

Due Date: 19 April 2021 Monday 11:59pm Anywhere on Earth

<u>Submission Instructions:</u> Save your write-up to a PDF document titled ***"YourLastName-CS2820-HW3.pdf"*** and e-mail it to your TA at pinar_demetci@brown.edu along with the supplementary material asked for (e.g. code), if any, by the deadline indicated above. Supplementary materials should also be similarly titled (e.g. "YourLastName-CS2820-HW3-Code.py") or included in your write-up.

Important Note

Given the short time we have for the rest of the semester and the workload from the final projects, we decided to follow a different route for this assignment. You <u>do not need to</u> answer all the questions.
Each question is marked with a number of points so just choose whichever ones you want to answer to get full 100 points. If you attempt to solve extra questions beyond that, you will get an extra credit up to 20 points.

1 Haplotype Phasing: Clark and EM Algorithms (50 Points)

1. Phase the following set of genotypes, following the steps of the Clark Algorithm. Write out the steps as you phase the genotypes and <u>show your steps</u>:

$$\begin{array}{cccccc} g_1 & 1 & 0 & 0 & 1 & 1 \\ g_2 & 1 & 2 & 0 & 2 & 1 \\ g_3 & 0 & 2 & 1 & 1 & 1 \\ g_4 & 1 & 2 & 1 & 0 & 2 \\ g_5 & 1 & 0 & 1 & 2 & 0 \\ g_6 & 0 & 1 & 0 & 0 & 0 \end{array}$$

2. What can you say about the importance of the order of these steps? How would the phase solutions change with different orders of applying these rules?

3. Write out a pseudocode for the EM algorithm for haplotype phasing. The main part to focus on is the calculation of the expected probability of a haplotype and making updates accordingly.

4. What are the strengths and weaknesses of these two approaches to haplotype phasing?

5. In your answer to the question above, you might have noticed that the advantages and shortcomings of Clark and EM algorithms are somewhat complementary. Can you come up with ideas to modify these methods and create a new algorithm that will combine the strengths of the two algorithms and improve the weaknesses? (Descriptions would suffice, no need for a pseudocode although we would not discourage you from it if you wanted to write one).

2 Implementing an MCMC algorithm (50 Points)

Please **<u>choose one</u>** of the following two questions. Implement a Metropolis-Hastings algorithm to solve it. Write up your results and turn in your code:

1. Sample from a Gaussian mixture model with two components of $N(0, 1)$ and $N(3, 1)$. Run the Markov chain $n = 50, 100, 500, 1000, 10000$ times and plot the histogram of points for each n. Overlay the true density function on the histogram.

OR

2. Crack the following code using the Metropolis-Hastings algorithm (tell us where it came from and how many words it has). What kind of encryption scheme is used? What are the encryption and decryption keys? Here is an online example/walk-through that will make your implementation easier: https://mlwhiz.com/blog/2015/08/21/mcmc_algorithm_cryptography/

```
kHrkn Bqqzonzg pbd gwn endg rq gzlnd zg pbd gwn pradg rq gzlnd
zg pbd gwn btn rq pzdkrl zg pbd gwn btn rq qrrizdwondd zg pbd
gwn nurhw rq eniznq zg pbd gwn nurhw rq zohankjizgv zg pbd gwn
dnbdro rq iztwg zg pbd gwn dnbdro rq kbafondd zg pbd gwn duazot
rq wrun zg pbd gwn pzogna rq kndubza pn wbk nmnavgwzot enqran jd
pn wbk orgwzot enqran jd pn pnan bii trzot kzanhg gr wnbmno pn
pnan bii trzot kzanhg gwn rgwna pbv zo dwrag gwn unazrk pbd dr
qba izfn gwn uandnog unazrk gwbg drln rq zgd orzdzndg bjgwrazgznd
zodzdgnk ro zgd enzot anhnzmnk qra trrk ra qra nmzi zo gwn
djunaibgzmn kntann rq hrlubazdro roiv gwnan pnan b fzot pzgw b
ibatn cbp bok b xjnno pzgw b uibzo qbhn ro gwn gwaron rq notibok
gwnan pnan b fzot pzgw b ibatn cbp bok b xjnno pzgw b qbza qbhn
ro gwn gwaron rq qabohn zo ergw hrjogaznd zg pbd hinbana gwbo
havdgbi gr gwn irakd rq gwn dgbgn uandnamnd rq irbmnd bok qzdwnd
gwbg gwzotd zo tnonabi pnan dnggink qra nmna zg pbd gwn vnba rq
rja irak ron gwrjdbok dnmno wjokank bok dnmnogvqzmn duazagjbi
anmnibgzrod pnan hrohnknk gr notibok bg gwbg qbmrjank unazrk bd
bg gwzd lad drjgwhrgg wbk anhnogiv bggbzonk wna qzmnbokgpnogzngw
einddnk ezagwkbv rq pwrl b uaruwngzh uazmbgn zo gwn izqn tjbakd
wbk wnabiknk gwn djeizln buunbabohn ev boorjohzot gwbg
baabotnlnogd pnan lbkn qra gwn dpbiirpzot ju rq irokro bok
pndglzodgna nmno gwn hrhfibon twrdg wbk enno ibzk roiv b arjok
kryno rq vnbad bqgna abuuzot rjg zgd lnddbtnd bd gwn duazagd rq
gwzd mnav vnba ibdg ubdg djunaobgjabiiv knqzhznog zo raztzobizgv
abuunk rjg gwnzad lnan lnddbtnd zo gwn nbagwiv rakna rq nmnogd wbk
ibgniv hrln gr gwn notizdw harpo bok unruin qarl b hrotandd rq
eazgzdw djecnhgd zo blnazhb pwzhw dgabotn gr anibgn wbmn uarmnk
lran zluragbog gr gwn wjlbo abhn gwbo bov hrlljozhbgzrod vng
anhnzmnk gwarjtw bov rq gwn hwzhfnod rq gwn hrhfibon earrk qabohn
indd qbmrjank ro gwn pwrin bd gr lbggnad duzazgjbi gwbo wna dzdgna
rq gwn dwznik bok gazknog ariink pzgw nshnnkzot dlrrgwondd krpo
wzii lbfzot ubuna lronv bok dunokzot zg jokna gwn tjzkbohn rq wna
hwazdgzbo ubdgrad dwn nognagbzonk wnadniq endzknd pzgw djhw wjlbon
bhwznmnlnogd bd dnognohzot b vrjgw gr wbmn wzd wbokd hjg rqq wzd
grotjn grao rjg pzgw uzohnad bok wzd erkv ejaonk bizmn enhbjdn wn
wbk org fonnink krpo zo gwn abzo gr kr wrorja gr b kzagv uarhnddzro
rq lrofd pwzhw ubddnk pzgwzo wzd mznp bg b kzdgbohn rq drln qzqgv
ra dzsgv vbakd zg zd izfniv norjtw gwbg arrgnk zo gwn prrkd rq
qabohn bok orapbv gwnan pnan tarpzot gannd pwno gwbg djqqnana pbd
ujg gr knbgw bianbkv lbafnk ev gwn prrklbo qbgn gr hrln krpo bok
en dbpo zogr erbakd gr lbfn b hnagbzo lrmbein qablnpraf pzgw b
dbhf bok b fozqn zo zg gnaazein zo wzdgrav zg zd izfniv norjtw
gwbg zo gwn arjtw rjgwrjdnd rq drln gziinad rq gwn wnbmv ibokd
bkcbhnog gr ubazd gwnan pnan dwnignank qarl gwn pnbgwna gwbg mnav
```

kbv ajkn hbagd endubggnank pzgw ajdgzh lzan dojqqnk berjg ev uztd bok arrdgnk zo ev urjigav pwzhw gwn qbalna knbgw wbk bianbkv dng bubag gr en wzd gjleazid rq gwn anmrijgzro ejg gwbg prrklbo bok gwbg qbalna gwrjtw gwnv praf johnbdzotiv praf dzinogiv bok or ron wnbak gwnl bd gwnv pnog berjg pzgw ljqqink ganbk gwn abgwna qrabdljhw bd gr nognagbzo bov djduzhzro gwbg gwnv pnan bpbfn pbd gr en bgwnzdgzhbi bok gabzgrarjd zo notibok gwnan pbd dhbahniv bo blrjog rq rakna bok uargnhgzro gr cjdgzqv ljhw obgzrobi erbdgzot kbazot ejatibaznd ev balnk lno bok wztwpbv areenaznd grrf uibhn zo gwn hbuzgbi zgdniq nmnav oztwg qblziznd pnan ujeizhiv hbjgzronk org gr tr rjg rq grpo pzgwrjg anlrmzot gwnza qjaozgjan gr juwridgnanad pbanwrjdnd qra dnhjazgv gwn wztwpbvlbo zo gwn kbaf pbd b hzgv gabkndlbo zo gwn iztwg bok enzot anhrtozdnk bok hwbiinotnk ev wzd qniirpgabkndlbo pwrl wn dgruunk zo wzd hwbabhgna rq gwn hbugbzo tbiibogiv dwrg wzl gwarjtw gwn wnbk bok arkn bpbv gwn lbzi pbd pbvibzk ev dnmno areenad bok gwn tjbak dwrg gwann knbk bok gwno trg dwrg knbk wzldniq ev gwn rgwna qrja zo hrodnxjnohn rq gwn qbzijan rq wzd blljozgzro bqgna pwzhw gwn lbzi pbd areenk zo unbhn gwbg lbtozqzhnog urgnogbgn gwn irak lbvra rq irokro pbd lbkn gr dgbok bok knizmna ro gjaowbl tanno ev ron wztwpbvlbo pwr kndurzink gwn ziijdgazrjd hanbgjan zo dztwg rq bii wzd angzojn uazdronad zo irokro tbrid qrjtwg ebggind pzgw gwnza gjaofnvd bok gwn lbcndgv rq gwn ibp qzank eijoknaejddnd zo blrot gwnl irbknk pzgw arjokd rq dwrg bok ebii gwznmnd dozuunk rqq kzblrok harddnd qarl gwn onhfd rq orein irakd bg hrjag kabpzotarrld ljdfngnnad pnog zogr dg tzindd gr dnbahw qra hrogabebok trrkd bok gwn lre qzank ro gwn ljdfngnnad bok gwn ljdfngnnad qzank ro gwn lre bok orerkv gwrjtwg bov rq gwndn rhhjaanohnd ljhw rjg rq gwn hrllro pbv zo gwn lzkdg rq gwnl gwn wbotlbo nmna ejdv bok nmna pradn gwbo jdnindd pbd zo hrodgbog anxjzdzgzro orp dgazotzot ju irot arpd rq lzdhniibonrjd hazlzobid orp wbotzot b wrjdneanbfna ro dbgjakbv pwr wbk enno gbfno ro gjndkbv orp ejaozot unruin zo gwn wbok bg onptbgn ev gwn kryno bok orp ejaozot ubluwingd bg gwn krra rq pndglzodgna wbii grkbv gbfzot gwn izqn rq bo bgarhzrjd ljaknana bok grlraarp rq b panghwnk uziqnana pwr wbk areenk b qbalnad erv rq dzsunohn bii gwndn gwzotd bok b gwrjdbok izfn gwnl hbln gr ubdd zo bok hirdn juro gwn knba rik vnba ron gwrjdbok dnmno wjokank bok dnmnogv qzmn nomzaronk ev gwnl pwzin gwn prrklbo bok gwn qbalna prafnk jownnknk gwrdn gpr rq gwn ibatn cbpd bok gwrdn rgwna gpr rq gwn uibzo bok gwn qbza qbhnd gark pzgw dgza norjtw bok hbaaznk gwnza kzmzon aztwgd pzgw b wztw wbok gwjd kzk gwn vnba ron gwrjdbok dnmno wjokank bok dnmnogv qzmn hrokjhg gwnza tanbgonddnd bok lvazbkd rq dlbii hanbgjand gwn hanbgjand rq gwzd hwarozhin blrot gwn andg birot gwn arbkd gwbg ibv enqran gwnl

3 GWAS Practice (50 Points — Courtesy of Daniel Ben-Isvy)

In this question, you will create a simulated dataset of genotypes and phenotypes and run a small GWAS pipeline, to select for genetic loci associated with the simulated phenotypes. GWAS simulations allow researchers to evaluate the performance of their methods under different models of population evolution and disease incidence.

3.1 Forward Simulation

SLiM is an evolutionary simulation framework that can be used to perform forward simulations of populations through time. Download SLiM from the website here. Take a look at the Chapter 4.1 of the SLiM manual, which includes an example Eidos script for neutral simulation. Following the example, Write an Eidos script to simulate the neutral evolution of a population of 10,000 diploid individuals over 20,000 generations. Each haplotype in the population should consist of one chromosome of length 1 megabase (Mb). Neutral mutations should arise in the population at a uniform rate of 1.1×10^{-8} mutations per base pair per generation, and recombinations should arise in the population at a uniform rate of 2.2×10^{-8} breakpoints per base pair per generation. After the population has evolved for 20,000 generations, use `outputFull()` to write all of the information about the final population to a text file. Please simulate two populations. Turn in your script or add it to your write-up. Chapters 4.2.1 and 26.1.1 of the SLiM manual contain details on the `outputFull()` method.

3.2 Case/Control Modeling

Consider an additive disease model with disease prevalence $p = 0.1$, disease allele frequency $q = 0.2$. For the first population, consider a relative risk of $GRR(Aa) = 1.1$ for the heterozygous genotype at the disease locus and for the second population, $GRR(Aa) = 1.5$. Simulate binary phenotypes, following the algorithm presented by Daniel Ben-Isvy in class for sampling a simulated case/control cohort from a population ($\epsilon = 0.005$). Turn in your simulations in a text file.

3.3 Variable Selection and Evaluation of GWAS Test Statistics

For this subquestion, **choose one** of the two paths below:

1. Use the `R` package `varbvs` (developed by Dr. Matthew Stephens' lab at University of Chicage) to compute the association probability for each locus, and create a scatter plot of these and highlight the disease loci in the plot. What can you say about the results?

   To do this, if you don't already have `R` on your computer, first install `R`. You might also want to install `RStudio`, which is a popular IDE to work with `R`. Then, install `varbvs` by running `install.packages(''varbvs'')`. Now you can run `varbvs` on your simulated genotype and phenotype data to find probability of association between each genetic locus and the phenotype by executing `results = varbvs(genotype, NULL, phenotype)`. Make sure your `genotype` is in a matrix format, where the rows correspond to individuals and the columns correspond to disease loci and your `phenotype` is in a vector format. You can access the association probabilities by typing `pips = results$pip`. `PIPs` here stand for "posterior inclusion probabilities". `varbvs` uses a variational EM procedure to optimize a sparse Bayesian regression function

   $$y = [\gamma N(0,1) + (1-\gamma)\delta_0]X + b$$

   where y is the phenotype, X is the genotype, and b is the residuals. The effect size of each genotype is denoted by $\gamma N(0,1) + (1-\gamma)\delta_0$, where γ is a Bernoulli variable, which takes the value of 1 with a probability of π and 0 with a probability of $1-\pi$. So this means that with a probability of π, genomic locus has an effect on the phenotype modeled by the standard normal distribution $N(0,1)$ and has no effect (δ_0 means a point mass at 0) with a probability of $1-\pi$. Inferring this probability π gives us the probability of association for a given genomic locus.

   OR

2. Compute the χ^2 test statistic for each mutated position in the cohort. You can drop any mutated positions that have expected counts of zero in any cell of the contingency table. Visualize the data on a scatter plot showing χ^2 test statistics on the y-axis and genomic positions on the x-axis. Plot the test statistic for the disease locus in a different color from the rest of the points.

4 Application of Clark & EM Algorithms (25 Questions)

1. Download the `clark.jar` and `em.jar` files from the HW page of the course website, as well as `genotypes.txt` file. Run the two algorithms on these genotypes to phase them using `java -jar clark.jar` or `java -jar em.jar`. (will be uploaded on the course website by the end of April 8, Thursday).

2. Comment on the similarities and differences you observe in the haplotypes inferred by the two methods. Which one seems more conservative? Did the Clark algorithm yield any orphan genotypes?

3. These genotypes were created by simulating haplotype data using the `ms` program, which was developed by Richard R. Hudson from University of Chicago. The `ms` program simulates genetic variation (in particular, SNP) data for randomly sampled haplotypes from a population, generated under Wright-Fisher neutral model (finite population size, discrete generations, and infinite sites model for variation). If you were curious and wanted to learn more about this program, here is the publication by Hudson (you do not need to read it to complete this assignment): Generating samples under a Wright–Fisher neutral model of genetic variation
The haplotypes used to simulate these genotypes are in `haplotypes.txt` file. What can you say about how the haplotypes inferred by the two methods compare to these?

4. Simulate a small dataset of haplotypes (10 samples with 5 sites in each) using the `ms` program yourself and and phase them to create 5 genotypes. To do this, download `ms.tar.gz` from: https://uchicago.app.box.com/s/l3e5uf13tikfjm7e1il1eujitlsjdx13. Ignore `msHOT` files (this is an extension of the ms program, which we will not deal with). Unzip `ms.tar.gz`. You should not have a folder named "`msdir`". In your terminal, go to this directory and type

```
gcc -o ms ms.c streec.c rand2.c -lm
```

. Now, you should be able to simulate data by typing in your command line:

```
./ms <nsam> <nreps> -t <theta>
```

Read the `msdoc.pdf` file, "The basic command line" section for the description of inputs and example runs (gene trees, crossing over and gene conversion, spatial structure and migration, exponentially growing or shrinking population etc parts are not relevant for our application since we just want you to run a basic case. However, you'll see that you can take into account a bunch of different factors in your simulations if you wanted to).

5 Reading & Conceptual Questions (25 Questions)

Take a look at the following papers. **Choose whether you want to answer the questions on haplotype phasing or GWAS** and answer **only** those:

- **(Haplotype phasing)**—Haplotype Phasing: Existing Methods and New Developments [This is a review on haplotype phasing methods from 2012.]
  https://www.nature.com/articles/nrg3054

  OR

- **(GWAS)**— Benefits and limitations of genome-wide association studies
  https://www.nature.com/articles/s41576-019-0127-1
- **(GWAS)**— Rare and common variants: twenty arguments
  https://www.nature.com/articles/nrg3118

Questions:

1. **(on haplotype phasing, paper 1):** In our GWAS chapter, we briefly talked about pros and cons of using genotype vs haplotype data (" to phase or not to phase in GWAS"). The review paper above mentions a case that demonstrates how the use of haplotype data can be beneficial for GWAS. Describe it (page 3).
2. **(on haplotype phasing, paper 1):** What are some constraints with respect to the genomic variants considered in haplotype phasing?
3. **(on haplotype phasing, paper 1):** Without ground-truth data on haplotypes, it can be difficult to measure the performance and accuracy of haplotype phasing algorithms. What are some metrics mentioned in the review paper?

   OR

4. **(on GWAS, paper 2):** Skim the paper and give a brief description of a few criticisms and limitations of GWAS discussed in this paper.
5. **(on GWAS, paper 3):** What is your take-away from the discussion of strengths and weaknesses of the two models discussed in the paper?