CS168 Practice Final

(Do not turn this page until you are instructed to do so!)

1. **Miscellaneous short answer.** (65 points)

   (a) (5 points) Recall the Count-Min Sketch data structure: we have m different "counters" $c_1, \ldots, c_m$ that are all initialized to 0, and we have a set of hash functions $h_1, \ldots, h_k$, that map items to the set $\{1, \ldots, m\}$. We observe a stream of numbers $x_1, x_2, \ldots$, and upon seeing the ith number, x_i, we increment each of $c_{h_1(x_i)}, \ldots, c_{h_k(x_i)}$. We also discussed the "conservative" version of this update rule, where instead of incrementing each of the k counters for each update, we only increment the smallest counters, namely the counters $c_{h_j(x_i)}$ for which $c_{h_j(x_i)} = \min_{\ell \in \{1,\ldots,k\}} c_{h_\ell(x_i)}$. In two sentences, explain the benefits of this "conservative" update rule, in contrast to the original update rule.

   (b) (5 points) What is the main problem that consistent hashing solves? Explain your answer in one to two sentences. [Recall the following sketch of the high-level idea of consistent hashing: there is a set S of m servers and a set O of n objects. We use a random hash function h to map all of S and O to a common set of number $\{1, 2, \ldots, N\}$ and then map each object $o \in O$ to the server $s \in S$ whose hash $h(s)$ is "closest" to the hash of the object, $h(o)$.]

(c) (5 points) Consider applying the Johnson-Lindenstrauss transform (i.e. projecting all points onto a set of randomly chosen directions) to a set of n points. Will the projected points have the property that for each of the $\binom{n}{3}$ subsets of 3 points, the area of the triangle spanned by those three points in the original space will be close to the area of the triangle spanned by the projections of the three points? Justify your answer in at most two sentences.

(d) (5 points) How can you select a d-dimensional vector uniformly at random from the ℓ_2 unit ball (namely, uniformly from the set of d-dimensional vectors that have Euclidean distance exactly 1 from the origin)? Describe the algorithm in at most one sentence, no justification required.

(e) (5 points) You go to the Great Barrier Reef and see many types of fish. After ten minutes of swimming, you make a tally of the fish you have seen:

Type of Fish	mackerel	parrotfish	clownfish	goby	grouper	angelfish	Wrasse
# Times Seen	5	3	2	2	1	1	1

Give an estimate of the probability that the next fish you see is a new type that you have not already seen, assuming that each fish you observe is an independent sample drawn from the distribution of fish in the reef? Show your calculation and provide at most one sentence of explanation—this can simply be a reference to the relevant result from lecture.

(f) (5 points) Consider the vector $v \in \mathbb{R}^{1024}$ that is the zero vector, except has ones in the 27th and 29th positions. Let $F(v) \in \mathbb{C}^{1024}$ denote the Fourier transform of v.

i. Will any of the indices of $F(v)$ be zero? If so, which one(s)? Provide a one-sentence justification.

ii. If v is a length 1023 vector that is identically zero, except has ones in the 27th and 29th positions, will any of the indices of $F(v)$ be zero? Provide a one-sentence justification.

(g) (5 points)

i. In one, clear sentence, explain what it means for a model to "generalize".

ii. Suppose you are given labeled datapoints (x, y) where $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$, and you would like to learn a cubic function of x that predicts y; roughly how many datapoints would we expect to need to ensure that the model you learn will generalize? Explain with one additional sentence.

(h) (5 points) Suppose you are arguing with a friend about whether its better to have a matrix of data, versus a tensor of data. 1) Describe one concrete and technical/rigorous reason why it can be better to work with matrices. 2) Describe one concrete and technical/rigorous reason why it can be better to work with a tensor.

(i) (5 points) Given an unweighted graph, the Laplacian matrix is defined as $L = D - A$ where D is the diagonal matrix of degrees, and A is the adjacency matrix. Explain 1) how, and 2) why the second-smallest eigenvector of L can be used to find a partition of the graph into two parts, such that there are relatively few edges crossing from one part to the other.

(j) (5 points) [This material will be from lecture 18] Farmer Leland owns a garden and a field, and is planning to grow some vegetables and hay. It costs \$5/acre to grow vegetables in the garden, but \$10/acre to grow it in the field. It costs \$3/acre to grow hay in the garden, and only \$1/acre to grow it in the field. Leland needs to grow at least 3 acres of vegetables to feed the family, and everything else will be sold at market for \$13/acre for vegetables and \$5/acre for hay. The garden has 7 acres, and the field has 100 acres. Formulate the problem of maximizing the amount of money that Farmer Leland can make as a Linear Program. (No need to solve it, just write the linear program. Note, we are NOT looking for python or matlab code.)

(k) (5 points) Which of the following sets are guaranteed to be convex? You do not need to provide a proof, just circle the convex sets:

i. the feasible region of a linear program
ii. the subset of $\mathbb{R}^d$ consisting of points whose ℓ_3 norm is at most 1
iii. the set of $n \times n$ rank-2 matrices (viewed as a subset of $\mathbb{R}^{n^2}$)
iv. the intersection of a collection of convex sets
v. the union of a collection of convex sets

2. **Singular Value Decomposition** In this question, we will examine SVD in the context of exploratory data analysis. Assume that the Senate consists of 100 senators and that they considered 1,000 bills during a given time period, with each senator voting either "Yes" or "No" on each of the 1,000 bills. We can represent this data as a 100×1000 matrix M, where the entry $M_{i,j}$ is 1 if the ith senator voted "Yes" on the jth bill, and is 0 otherwise.

(a) (5 points) To begin, lets assume that each senator is affiliated with either party A or party B, split roughly 50/50, and that all members of a party cast the same vote ("Yes/No") on a bill. Additionally, assume that each of the bills belongs to exactly one of the following three types: 1)

supported by both parties, 2) supported only by party A, 3) supported only by party B. Under these assumptions, what is the rank of matrix M? Justify your answer with at most two sentences.

(b) For the remainder of this question, lets relax the above assumptions, and instead, just assume that, on most bills, members of the same party generally vote similarly. Consider the voting matrix M, and its singular value decomposition $M = U\Sigma V^T$.

i. (3 points) Would the top few left singular vectors (columns of U) have any natural interpretations in this case? Explain your answer in a sentence or two.

ii. (3 points) Would the top few right singular vectors (columns of V) have any natural interpretations in this case? Explain your answer in a sentence or two.

iii. (5 points) Given the SVD decomposition $M = U\Sigma V^T$, how can you compute a good rank r approximation of M? In what formal sense is the approximation you describe the "best" rank r approximation of M?

iv. (3 points) Roughly what would the best rank-1 approximation for M correspond to? (For example. what is a plausible interpretation of the row and column vector that define this rank-1 matrix?)

v. (3 points) Roughly what would you expect the best rank-2 approximation for M to correspond to, and how would this differ from the matrix M?

3. **Sampling and Estimation.** Suppose you are conducting an expensive study to determine what fraction of the population has a certain common genetic mutation. In the following parts, we consider several possible alternate scenarios.

   (a) (5 points) Suppose you obtain genetic sequences for 10,000 randomly chosen people, of which 2,480 have the mutation. Roughly what do you expect the percent error in the outcome of study to be? Explain your answer with at most two sentences. (You might find it helpful to refer to Chebyshev's inequality, which states that for a random variable X and any $c > 0$, $\Pr\left[|X - E[X]| > c\sqrt{Var[X]}\right] \leq 1/c^2$.)

   (b) (5 points) Suppose you *know* that exactly 50% of the population are from demographic A (based on census data, for example), and the rest are not. You have a hunch that most people in demographic A have the mutation and most people outside of demographic A do not have the mutation. How could you design a study of 10, 000 people so that 1) the expectation of the answer you report is the true percentage of the population with the mutation, and 2) if your hunch is correct, you get a better percentage error than you got in the previous part? Explain your answer with at most two sentences.

   (c) (5 points) Suppose you *know* that exactly 50% of the population are from demographic A , and the rest are not. You have a hunch that roughly half of the people in demographic A have the mutation, and only a quarter of the people outside of demographic A have the mutation. You have enough funding to do a study on 10, 000 people—how can you design your study to get the smallest percentage error possible, assuming your hunch is correct? Explain your answer with at most two sentences.

4. **Max 3-Cut with MCMC.** Suppose we are given an undirected weighted graph $G = (V, E)$ and wish to partition the graph into three parts so as to maximize the sum of the weights of the edges that go between different sets. This problem is NP-hard, in the worst case. Nevertheless, in this question we will design a Markov Chain Monte Carlo search heuristic for finding such a set, which would work quite well in many settings.

   Assume that the number of vertices is $|V| = n$, and that each edge $(u, v) \in E$ has a non-negative weight $w_{(u,v)}$. The *weight* of a partition of V into sets V_1, V_2, V_3 is defined to be the sum of the weights of each edge that has endpoints in different sets:

$$w(V_1, V_2, V_3) = \sum_{(u,v)\in E \text{ s.t. } u\in V_i \text{ and } v\in V_j \text{ with } i\neq j} w_{(u,v)}$$

   Suppose you attempt to find a 3-partition (V_1, V_2, V_3) of maximum weight $w(V_1, V_2, V_3)$ via an MCMC approach:

   (a) (3 points) How many states (as a function of n) are in this Markov Chain? Feel free to just give your answer in big-Oh notation, accurate up to a constant factor.

   (b) (5 points) At each iteration of MCMC, we need to propose a new partition based on the current partition. Among the following two proposal schemes, which scheme is better? Justify your answer with a brief explanation:

      i. Randomly select one node, and randomly put it in one of the other two sets.

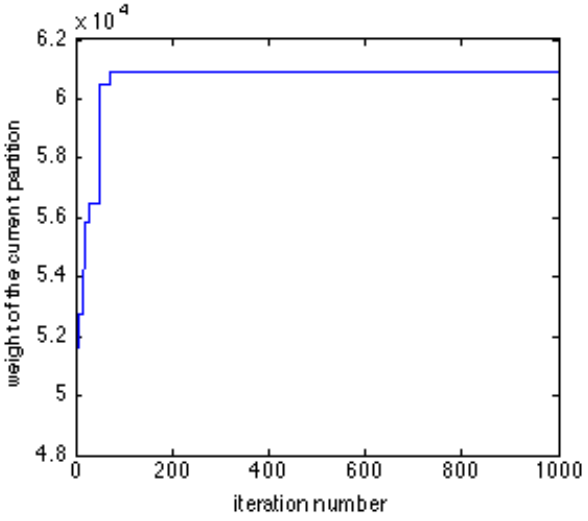
      ii. Select two random nodes that are in different sets, and swap their sets.

   (c) (5 points) The transition function of our MCMC algorithm will be parameterized in terms of some "temperature" T. Complete the following pseudo code of the MCMC algorithm in the gap provided. [Hint: you might consider using the quantity $q = e^{\Delta/T}$ or $q = e^{-\Delta/T}$, where Δ is as defined in the pseudo code below.]

```
partition ← random partition of the N nodes
best ← partition
for i = 1, 2, ..., MAXITER do
    partition_new ← new proposed partition based on the current partition
    Δ = w(partition_new) - w(partition)

    if w(partition) > w(best) then
        best ← partition
    end if
end for
```

(d) (5 points) After setting the total number of iterations MAXITER and the temperature, T, you run the MCMC algorithm. You plot the iteration number against the weight of the current partition and see the following plot:



If you want to improve the MCMC search to improve the chances of finding a better partition, which of the following is a reasonable thing to try next? Circle one of the options and justify that answer with at most 2 sentences.

i. Increase temperature T.

ii. Decrease temperature T.

iii. Increase the number of iterations.

iv. Decrease the number of iterations.

5. Suppose that when you play an audio signal on your computer, you know that the defects in your computer's speakers end up adding some echoes and distortion to the signal. Specifically, the sound that comes out of the speakers is the convolution of the true signal, s with a filter f, so you end up hearing the distorted signal $s * f$.

(a) (5 points) First, suppose you would like to learn f; what signal, s, can you design such that by recording what happens when you try to play signal s, you end up hearing f? In other words, what signal (i.e. vector) s has the property that $s * f = f$? (No explanation necessary.)

(b) (5 points) Now that you know f, what signal should you ask the computer to play so that you hear true signal s. Namely, for what signal r is it the case that $s = r * f$? (No explanation necessary, just write the formula for r in terms of s and f.)

(c) (7 points) In one or two sentences, explain what could go wrong with the above attempt to 'fix' the echoes and distortion in your speaker? Give an example of a type of distortion, f, for which the 'fix' of the previous part would be expected to work well, and an example of a type of distortion for which the above approach would not work well. Provide a one or two sentence justification for your answers.

6. **The Algorithmic Toolbox.** For each of the following scenarios, select the technique/tool from the list that is best suited to the problem. Note that some tools/techniques might be matched with several scenarios, and some tools might not be matched with any scenarios. For each question except (c), identify ONE tool/technique, and **provide a one-sentence explanation of how or why the technique should be applied.** Part (c) asks you to identify TWO tools/techniques.

   Reminder of tools/techniques: Consistent Hashing, Count-Min-Sketch, Min-Hash, Locality Sensitive Hashing, Dimension-Reduction, k-d Trees and Nearest Neighbor Search, Importance Sampling, Markov Chain Monte Carlo, PCA, SVD and low-rank matrix approximation, Tensor Methods, Spectral Graph Theory, Fourier Analysis/Convolution, Linear/Convex Programming, Differential Privacy.

   (a) (3 points) You are pursuing a PhD in biology, and are trying to understand the potential consequences that global warming might have on plant diversity. You conduct a large experiment as follows: you will grow a few hundred different species of plants, and for each one, you grow specimens in 20 different atmospheric settings, for example a 'high CO2' setting, a 'high methane' setting, a 'elevated ozone' setting, etc. Additionally, for each of these plant/atmosphere combinations, you do two copies of the setup, one where the plants are grown in 'normal' temperature conditions, and one under 'warmer' temperature conditions. How might you analyze the growth data that you collect, to try to extract some high-level insights into the broad types of effects of the different atmospheric conditions/temperature combinations?

(b) (3 points) You are leading a team to combat the recent surge of drug-resistant-bacteria and want to apply computational techniques to help identify some likely compounds that might form the core of new antibiotics. You make a large matrix whose rows and columns are indexed by chemical compounds, and the (i, j)-entry is the strength of the interaction between compounds i and j. You have tested a number of pairs of compounds, and would like to fill in the missing values in the matrix with estimates of their interaction strength.

(c) (6 points) Based on the success of the previous project, you are promoted to head data scientist for the CDC (Center for Disease Control), where you continue your work on antibiotic resistant bacteria. You discover that many of the new antibiotic resistant bacteria originate from one reasonably small geographic area that contains hundreds of chicken farms. Currently, all the chicken farms use *all* 6 commonly available antibiotics. With the hopes of being able to contain the evolution and spread of new antibiotic resistant strains of bacteria, you want to assign 1 of the 6 antibiotics to each of the farms, in such a way that nearly every farm uses an antibiotic that is different from the antibiotics used by the neighboring farms. Give TWO different approaches to deciding this assignment of antibiotics to farms.

(d) (3 points) You work for Teach for America, and you suspect that there are some curious relationships between the number of books that students have read, their parents' ages, their parents' incomes, the amount of one-on-one time that they spend with adults, and their GPAs. What tool might you use to try to make sense of the relationships between these variables?

(e) (3 points) After your Teach for America service, you decide to start a company that searches through patents, looking for potential similar/duplicate patents (perhaps with the aim of offering expensive legal counsel to one of the parties). There are millions of patents to sift through—how might you tackle the computational challenge of this?

(f) (3 points) You are a scientist at NASA trying to design a proof-of-concept mission to extract gold from a nearby asteroid. The spaceship (which has relatively little computational power, because big batteries are heavy), must compute an accurate map of the density of the asteroid (to find which parts have the most gold...of course!), based on gravitational measurements that it will compile while orbiting the asteroid. What tool would you use to very efficiently make calculation that "inverts" the gravitational field to deduce the density map of the asteroid?