Final Exam

Electronic submission via Gradescope, due **11:59pm Wednesday 12/13**.

You must work alone. You may consult the textbook/course notes, and other online resources (wikipedia, etc.). You may not discuss any aspect of this test with other students (whether or not they are enrolled in the class). You may NOT troll the internet looking for problem-specific solutions, or post course-related questions on online forums, with the exception of our class Piazza forum.

Please answer the questions thoroughly yet succinctly. Points may be deducted for needlessly elaborate solutions, or for vagaries or poor style. As always, if there is a hole in your proof, fewer points will be deducted if you acknowledge that you are aware of this hole, versus trying to hide it. Good luck!

1. For each of the following questions, provide a one to three sentence explanation.

   (a) (3 points) Suppose I just purchased gradescope’s new “automatic grader” algorithm, which promises that it will grade each person’s final exam, and comes with the guarantee that, no matter the distribution from which students’ exams are drawn, the average number of grading errors that the program will make per exam is at most $\alpha$. Assuming the guarantee is true, I would like to say something about what fraction of students will end up with lots of grading errors. What can I do?

   We can apply Markov’s inequality, since the random variable representing the number of grading errors on a test is non-negative; this yields that for any value $\lambda$, the expected fraction of students with more than $\lambda$ grading errors is at most $\alpha/\lambda$. [For example, the probability of having more than $10\alpha$ errors is at most $1/10$.

   (b) (6 points) Describe three rather different approaches to showing that two distributions, $A$ and $B$ actually have the same distribution, or have very similar distributions, and indicate when each approach might be useful. Your three approaches could apply to different settings—for example, one approach could apply to the case where the distributions are continuous distributions over the reals, and are described to you via some complicated sums of other continuous random variables; another could apply to the case where the distributions $A$ and $B$ are described implicitly as the distributions resulting from applying some random process (e.g constructing a random graph via a strange process of drawing edges); perhaps you are given access to independent draws from $A$ and $B$. As long as you clearly describe three approaches, and indicate when that approach might be useful, you will get full credit.

   If we have some description of a pair of real-valued distributions, one approach to showing that they are the same is to show that their moment-generating-functions agree (this is a nice way of proving that a sum of Gaussians is Gaussian). If we are given a pair of
distributions defined via some strange processes, one approach is to define a joint distribution whose marginals agree with the two distributions, and argue that the probability that the two components of the joint are equal is close to 1. A third approach is the dual of this joint approach: try to directly prove that for any subset \( S \) of possible outcomes, the probability mass that the two distributions put on events in set \( S \) are similar or identical.

(c) (3 points) Suppose you want to show the existence of some nice combinatorial object (such as a good encoding/decoding scheme, or a graph with special properties, etc.) What is one general approach to showing that such an object exists?

One approach is via the ‘Probabilistic Method’: define a distribution \( D \) over objects, and show that an object selected according to \( D \) has a non-zero probability of being the desired “nice” object.

(d) (3 points) Consider a dataset of \( n \) genomes, each represented as a length \( m \gg n \) binary vector. You would like to compute all \( n^2 \) pairwise Hamming distances between the vectors, but you do not have time \( O(n^2m) \). Instead, you decide to approximate each of these pairwise distances, and would be happy if most of the distances are accurate up to 10% error. What could you do, and what is the amount of runtime it would take to compute these \( n^2 \) approximate distances? [Justify your answer, thought feel free to ignore constant factors.]

Consider first randomly (via Johnson-Lindenstrauss) project each vector to dimension \( O((\log n / 0.1^2) = O(\log n) \), and then explicitly calculate the pairwise Euclidean distances in this lower-dimensional space and square each computed distance, the runtime will be \( O(mn \log n + n^2 \log n) \), where the first term is the time it takes to compute the projection of the \( n \) vectors, and the second term is the time to compute (and square) the Euclidean distances in the projected space. The Johnson-Lindenstrauss result guarantees that, with high probability (for an appropriately chosen constant in the \( O(\log n) \) dimension), the Euclidean distances will all be preserved up to multiplicative error, say, 1%. To conclude, note that the Euclidean distance is the square-root of the Hamming distance, hence the recovered Euclidean distance, \( q \in (0.99\sqrt{d}, 1.01\sqrt{d}) \), and hence \( q^2 \in (0.99^2d, 1.01^2d) \). The claim now follows from observing that \( 0.99^2 \approx 0.98 > 0.9 \) and \( 1.01^2 \approx 1.02 < 1.1 \).

(e) (6 points) Suppose you have \( n \) jobs, that you are allocating uniformly at random to \( m \) processors (i.e. for each job, you pick a number uniformly at random from 1, \ldots, m and assign that job to the corresponding processor). Let \( X \) denote the number of processors that end up with more than 2 jobs. Describe (in high-level terms) two different approaches for proving that \( X \) will be tightly concentrated about its expectation. [Here, tightly concentrated means with inverse exponentially decreasing tail bounds, as in a Chernoff bound.]
One approach would be via ‘Poissonization’, another approach is via the Doob martingale and Azuma-Hoeffding tail bounds. In the Poissonization setting, we consider tossing $k \leftarrow \text{Poi}(n)$ balls, and hence the bin loads are independent and thus are amenable to Chernoff bounds. To connect that result of the Poissonized setting to the original setting, we use the fact that whp $|k - n| < \sqrt{n} \log n$, and the addition or removal of these “extra” balls can’t alter the result by more than this discrepancy. For the martingale approach, we let $Z_i$ denote the expected number of bins with more than 2 jobs after the first $i$ balls have been tossed, and the expectation is over the remaining $n - i$ balls. This martingale has a bounded difference (bounded by 1), so we can apply Azuma-Hoeffding tail bounds.

(f) (3 points) Suppose you have an enormous dataset of chord progressions found in pop music. You want to make a program that generates random (ideally novel) pop-sounding musical phrases. As an initial step, you first make a simple subroutine that takes a pair of phrases/sequences, and estimates their relative likelihoods of arising in pop music. (Such a subroutine could be based on the relative likelihoods in your dataset of all the 4-note subsequences, for example.) At a very high level, in one or two sentences, describe what techniques you might use to leverage this subroutine into a program that generates sequences according to some reasonable approximation of the distribution of pop-music phrases.

We can apply some version of the Metropolis MCMC algorithm (in this case, it will be close to the special case of Gibbs sampling), by starting from a random sequence, and considering altering small subsections, and accepting such a perturbation with probability related to the ratio of likelihoods of the original vs perturbed sequence.

(g) (3 points) Suppose there are $n \geq 3$ houses on your street, arranged in a line. At time $t = 0$, a soccer ball appears in one of their yards. Suppose that each day, whichever house has the soccer ball, they wake up each day and leave it with probability $p$, or with probability $1 - p$ kick it into a random adjacent yard (the two end houses only have 1 choice of adjacent yard, and all the other houses have 2 choices of which yard to kick it into). Assuming all houses have the same probability of inaction on a given day, $p \in (0, 1)$, what do you expect is the distribution of the ball’s location at day $t$ for some extremely large number $t$?

Because $p \in (0, 1)$, the location of the ball is a finite, irreducible, aperiodic Markov Chain, and hence has a unique stationary distribution. This stationary distribution will have equal probability $q$ on all the non-end houses, and probability $q/2$ on the two end houses. (Since the probabilities add to 1, this yields $q(n - 2) + 2q/2 = 1$ which implies that $q = 1/(n - 1)$.) To see that this is a stationary distribution (and hence is THE stationary distribution), suppose that at time $t$, we are at this distribution, and consider the distribution at time $t + 1$. For each non-end house, the probability is $qp + q(1 - p)/2 + q(1 - p)/2 = q$, where the first $qp$ is the probability we get from keeping the ball at this house, and each $q(1 - p)/2$ term is the probability that we get from each neighbor possibly kicking it into our yard. [Note that the $q(1 - p)/2$ term holds for both non-end neighbors, as well as the end houses—if our neighbor is an end
house, they kick it into our yard with probability \( (q/2)(1-p) \). For the end houses, a similar calculation yields \( (q/2)p + q(1-p)/2 = q/2 \), as desired.

(h) (3 points) In the above setting, if \( p = 0 \) (i.e. the ball gets kicked every day) describe the distribution of the ball’s location at time \( t \) for some extremely large number \( t \).

Suppose the houses are numbered \( 1, 2, \ldots, n \). If the ball starts at an even-numbered house, then for large \( t \), if \( t \) is even, the probability it is at an odd house is 0, and for even \( i \), the probability it is at house \( i \) is \( w \) for all \( i \) except for the two ‘end’ even houses (house \( i = 2 \) and house \( i = n \) or \( i = n − 1 \)—whichever is even) where the probability is \( w/2 \). Similarly, if \( t \) is odd, then the probability is 0 for even-indexed houses, and is equal to \( w' \) for all houses except the largest and smallest odd-indexed houses where the probability is \( w'/2 \). As in the previous part, we can solve for \( w \) or \( w' \) trivially. To see why these answers are correct, note that if we define the Markov Chain corresponding to only the ball’s locations at even [odd] timesteps, the chain has states corresponding to all the even-indexed [odd-indexed] houses, and corresponds to the chain of the previous part with \( p = 1/2 \), and hence the results of the previous part apply to this ‘sped-up-by-2’ chain.

2. Suppose there are \( n \) different classes, and \( m > 2n \) possible TAs. Each class needs at least 2 TAs, and no one should TA more than one class. Suppose each class submits a list of 20 “preferred” TAs, and it so happens that each TA appears in at most 3 lists.

(a) (6 points) Prove, via the Lovasz Local Lemma, that there exists an assignment of TAs to classes such that every class gets at least two of its preferred TAs. [Hint: Consider assigning each TA uniformly at random to one of the \( \leq 3 \) classes that prefer that TA. Be sure to formally define the “bad” events, and argue about the degree of the dependency graph.]

Consider assigning each TA uniformly at random to one of the \( \leq 3 \) classes that prefer that TA. For each class, define a “bad” event that fewer than 2 TAs are assigned to that class. A valid dependency graph will connect any two classes that have at least one ‘preferred’ TA in common. [This is valid because whether a bad event happens is determined only by the allocation of its preferred TAs, and the non-neighbor events do not involve any of those potential TAs.] For the LLL to hold, we just need to ensure that for each bad event, \( A \), \( \Pr[A] \leq \frac{1}{e(d+1)} \), where \( d = 2 \times 20 = 40 \) is an upper bound on the degree of the dependency graph—this expression comes from the fact that each of our 20 preferred TAs can be in at most 2 other class lists. \( \Pr[A] = (2/3)^{20} + 20(2/3)^{19}(1/3) < 0.004 < 0.008 < 1/(e(40 + 1)) \).

(b) (2 points) Describe an algorithm, whose runtime is polynomial in \( n \) and \( m \) that could be used to find a satisfactory assignment. (No proof necessary, just clearly describe the algorithm.)
We apply the constructive LLL algorithm: First pick an arbitrary allocation of TAs to a course that lists them. Then, while there exists an understaffed course, choose one such course and “re-sample” the allocation of all 20 of that class’s preferred TAs, where each of these preferred TAs is assigned a uniformly random course out of their \( \leq 3 \) course choices that list them.

3. Suppose the class has \( A \) CS students, \( B \) Statistics students, and \( C \) Engineering students. Suppose a uniformly random subset of \( k \) students ends up submitting a course evaluation. Let \( X \) denote the number of CS students who submitted evaluations.

(a) (3 points) What is \( \mathbb{E}[X] \) as a function of \( A, B, C, \) and \( k \)? [Prove your answer with at most one sentence.]

By linearity of expectation, \( \mathbb{E}[X] = k \frac{A}{A + B + C} \). To see why this is linearity of expectation, note that each of the \( k \) students is a uniformly random student (their joint distribution is NOT \( k \) i.i.d. draws, but the marginal of each of the \( k \) students' identities is an i.i.d. draw).

(b) (1 point) Suppose we iteratively select our random set of \( k \) students, by first choosing a random student to add to the set, then selecting a random student from the \( A + B + C - 1 \) remaining ones, and so on until we have selected \( k \) students. Prove, in one sentence, why the resulting set consists of a uniformly random subset of \( k \) students.

For each specific set of \( k \) students, the probability of drawing the set via this process is exactly \( k! \) times the product of \( 1/(A + B + C) \), \( 1/(A + B + C - 1) \), etc., which is the probability in the uniform distribution (i.e. it is identical for all sets).

(c) (3 points) Let \( X_i \) be the indicator random variable of whether the \( i \)th student chosen according to the above sequence is a CS student. Hence \( X = \sum_{i=1}^{k} X_i \). Are the \( X_i \)'s independent or dependent random variables? Prove your answer.

Dependent. For example, if the first student is a CS student, the probability \( X_2 \) is a CS student is smaller than if the first student was NOT a CS student.

(d) (6 points) Prove that \( \Pr[|X - \mathbb{E}[X]| > \lambda] \leq 2e^{-\frac{\lambda^2}{2}} \).

We will consider the Doob martingale and then apply the Azuma-Hoeffding tail bound. Let \( Z_i \) denote the expectation of \( X \), conditioned on \( X_1, \ldots, X_i \). We claim this martingale has differences bounded in magnitude by 1, since this conditional expectation at time \( i + 1 \) is maximized if \( X_i = 1 \) and is minimized if \( X_i = 0 \), and in either case, we can explicitly calculate what this expectation is via linearity of expectation (as in part a), and this difference is bounded in magnitude by 1. The result then follows from the Azuma-Hoeffding tail bound.
4. Consider a fair betting game, where you bet \( i \) dollars at time \( i \) (and either win \( i \) or lose \( i \) dollars, independently, with probability \( 1/2 \) each). Let \( X_i \in \{\pm1\} \) denote whether you win or lose in the \( i \)th round.

(a) (1 point) Let \( Y_i = \sum_{j=1}^{i} jX_j \) denote your net winnings through the first \( i \) rounds. Prove that \( \{Y_i\} \) is a martingale with respect to \( \{X_i\} \).

To prove that \( \{Y_i\} \) is a martingale wrt \( \{X_i\} \), note that \( \mathbb{E}[Y_{i+1}|X_1, \ldots, X_i] = Y_i + i(1/2) - i(1/2) = Y_i \).

(b) (5 points) Choose a function \( f(i) \) such that if you define \( Z_i = Y_i^2 + f(i) \), then the sequence \( \{Z_i\} \) is a martingale with respect to \( \{X_i\} \). [Hint: If you don’t know where to start, first figure out what \( f(1) \) should be so that \( \mathbb{E}[Y_i^2] + f(1) = 0 \), then figure out what \( f(2) \) should be, then \( f(3) \), etc. until you see the pattern, and then prove that it works.]

Consider setting \( Z_i = Y_i^2 - \sum_{j=1}^{i} j^2 \). To see that this is a martingale wrt the \( X_i \)’s, consider

\[
\mathbb{E}[Z_{i+1}|X_1, \ldots, X_i] = (1/2)(Y_i+i+1)^2+(1/2)(Y_i-(i+1))^2 - \sum_{j=1}^{i+1} j^2 = Y_i^2 + (i+1)^2 - \sum_{j=1}^{i+1} j^2
\]

\[
= Y_i^2 - \sum_{j=1}^{i} j^2 = Y_i;
\]

hence \( \{Z_i\} \) is a martingale as claimed.

(c) (5 points) Let \( T \) denote the (random variable) representing the first time when your net earnings are either more than \( a \) or less than \( -a \). What is \( \mathbb{E}[T] \) as a function of \( a \)? Its fine if your answer is only accurate up to a constant factor. [If you use the optional stopping theorem, be sure to argue why you can apply it to this setting.]

We will first prove an upper bound on \( \mathbb{E}[T] \), establishing that \( \mathbb{E}[T] = O(a^{2/3}) \). We will then show the matching lower bound, that \( \mathbb{E}[T] = \Omega(a^{2/3}) \), which together establish that \( \mathbb{E}[T] = \Theta(a^{2/3}) \).

**Upper bound:** The optional stopping theorem applies as \( T \) is bounded in magnitude by \( 2a \). The stopping theorem yields \( \mathbb{E}[Z_T] = \mathbb{E}[Z_0] = 0 \). Plugging in the definition of \( Z \) yields \( \mathbb{E}[Y_T^2] - \mathbb{E}[\sum_{j=1}^{T} j^2] = 0 \). Noting that, by definition, \( Y_T^2 \in [a^2, 9a^2] \), and \( \sum_{j=1}^{T} j^2 > T^3/3 \), yields that \( \mathbb{E}[T^3] \leq 27a^2 \), and hence \( \mathbb{E}[T] \leq 3a^{2/3} \). [If \( \mathbb{E}[T] > c \), then \( \mathbb{E}[T^3] > c^3 \ldots ]

**Lower bound:** There are a few ways of doing this. Here is one. First note that from our upper bound, by Markov’s inequality, the probability that \( T > 10a^{2/3} \) is less than \( 1/2 \).
5. In this problem we will consider two different models of wealth distribution. Suppose we are in a world with \( n \) people and 100n dollars. At time \( t = 0 \) there is some initial assignment, \( X_0 \), of dollars to people.

**Model A:** At each time \( t = 1, 2, 3, \ldots \), the allocation changes according to the following protocol: choose a person uniformly at random from the \( n \) players, and call them player \( i \); if \( i \) has no money, then nothing changes during that time-step; if \( i \) has at least one dollar, then select a player, \( j \), with probability proportional to the amount of money they currently have, and transfer one dollar from \( i \) to \( j \) (note that it is possible that \( i = j \)). For example, if at time \( t \), all players have $100 then at time \( t + 1 \), with probability \( 1/n \), all players have $100, and with probability \( 1 - 1/n \), one player has $101 one has $99 and the rest have $100.

**Model B:** At each time \( t = 1, 2, 3, \ldots \), the allocation changes according to the following protocol: choose a person uniformly at random from the \( n \) players, and call them player \( i \); then select a player, \( j \), with probability proportional to the amount of money they currently have, and transfer one dollar from \( j \) to \( i \) (it is possible that \( i = j \)). [Note that in this model, \( j \) pays \( i \), whereas in Model A, \( i \) pays \( j \)!!]  

(a) (2 points) Note that each model describes a Markov Chain over the “wealth allocation” of players. For each model, say whether the chain is periodic or aperiodic, and give one sentence of justification.

Both models are aperiodic, since for every state, it is possible to stay in that state.

(b) (2 points) For each model, say whether the chain is irreducible or not, and give one sentence of justification.

Model B is irreducible, since for any pair of allocations, there is a sequence of at most 100n transitions that each occur with positive probability that will get from the first state of the pair to the second state. Model A is not irreducible, since from the allocation where player 1 has all the money, one cannot access any other allocation, ever.
(c) (2 points) For each model, say whether or not the fundamental theorem of Markov chains applies, and give a one sentence justification.

The fundamental theorem applies to Model B, as it is (finite), irreducible, and aperiodic.

(d) (4 points) What is the stationary distribution of the Markov Chain described by Model B? (Prove your answer.) [Hint: Imagine that each of the dollars has a unique number, 1, \ldots, 100n, and think about the state of the system at time t as being a list of which dollars each person has. Consider the update protocol that first picks a uniformly random dollar \( d \), then picks a uniformly random person \( j \), and takes that dollar from whoever has it and gives it to player \( j \). This update scheme is one way of implementing Model B, no? Now, if you think of the states of this chain as nodes in a graph, the updates look kindof like a random walk on this graph….and you know all about the stationary distribution of random walks on a graph : )]

Consider the Markov Chain over the larger state space where we label each dollar 1, 2, \ldots, 100n, and keep track of who has which specific dollars. As suggested by the hint, Model B corresponds to choosing a uniformly random dollar, and giving it to a uniformly random player. For each state in this state space (i.e. for every partition of this set of 100n dollars to the \( n \) players), there are \( 100n \cdot n \) equally-weighted transitions, of which \( 100n \cdot (n - 1) \) yield a different state (in this larger state space where we keep track of the identities of the dollars). Since the corresponding chain is irreducible, this Markov Chain corresponds to a lazy random walk on this \( 100n(n - 1) \)-regular graph (where the probability of remaining in the same state is exactly \( 1/n \) at each timestep). The stationary distribution is hence the \textsc{Uniform} distribution over the states, assigning probability \( \frac{1}{(100n)^n} \) to each state, as each of the \( 100n \) dollars can be at each of \( n \) different player. [And, if you want to know the probability in the stationary distribution of the wealth allocation \( (a_1, \ldots, a_n) \) denoting the state where player \( i \) has \( a_i \) dollars, that will just be proportional to the number of different ways that this allocation can be realized, namely \( \frac{1}{(100n)^n} \cdot \binom{100n}{a_1} \cdot \binom{100n-a_1}{a_2} \cdot \cdots = \frac{100n!}{a_1!a_2!\ldots a_n!} \).

(e) (5 points) Prove that the mixing time of the chain corresponding to Model B is at most \( O(n \log n) \). [Hint: define a coupling, and also use the hint of the previous part.] You can do this part even if you skip the previous part.

We consider the Markov Chain over the larger state space as in the previous part, and note that the mixing time of this chain is at least the mixing time of the smaller state-space chain correspondingly to the wealth allocation, because that chain corresponds to grouping all the states that correspond to the same monetary allocation, and combining domain elements can only reduce the total variation distance between two states. The coupling here is easy: consider two copies of this chain, in the first, we pick a uniformly random dollar from the set \{1, \ldots, n\} and a uniformly random person from \{1, \ldots, n\} and give the dollar to that person. In the second, chain, we pick THE SAME dollar and THE SAME player. The chains will be coupled as soon as all the dollars have been
picked, which is simply the coupon collector problem with $100n$ coupons, and hence the expected time until coupling is bounded by $C = 100n \log(100n) = O(n \log n)$, and via Markov’s inequality, the probability that the chains haven’t coupled after $(2e)C$ time is at most $1/2e$, and hence this bounds the mixing time.

(f) (2 points bonus) As $t$ gets large, in Model A one player will eventually end up with all the money. What is the expected time until this happens? Feel free to just give the answer up to constant factors, though give a rigorous proof.

This is a cool problem, and I’ll keep the solution a secret in case I want to use it as a pset problem next year : )