1 Rules of the exam

Before beginning, please read carefully the following ground rules of the exam. Please e-mail me a pdf of the exam at michaelm@eecs.harvard.edu by noon on May 1, but earlier if you can/finish it earlier. Please contact me if you need to make other arrangements. Do not turn in the exam late.

You are expected to spend a time period of 72 consecutive hours on the exam. That is, when you decide to start the exam, you have 72 hours from that point to complete the exam. This is an honor code sort of thing. Don’t mess with it.

You are allowed to use your own class notes and the book. You are not allowed to use other sources, including other books or information available on the Web, for this exam. There is one exception: if you are looking for specific identities or tail bounds, you may look for them on the Web, but you must cite them. (For example, you can say:

$$\sum_{j=0}^{k} \binom{m}{j} \binom{n-m}{k-j} = \binom{n}{k},$$

as given in http://en.wikipedia.org/wiki/Binomial_coefficient.) Any detection of utilizing sources improperly will result in an immediate failing grade and a report to the Administrative Board. If you have any questions regarding this policy please ask.

You are not, under any circumstance, to talk with others about the exam before May 2.

In many cases, you may be asked to give a bound. When bounds are left unspecified, it is assumed that you will try to achieve the best bound possible (within reason; as usual, in theory, we often don’t mind being off by constant factors so much). If any question seems otherwise vague (or potentially incorrect!) please let me know as soon as possible.

You are to answer any five of the following. You may answer six, and I will take the best five scores. The exam is quite difficult; I suggest that focusing on five problems is probably better than trying six, but that is up to you. Partial credit will be given, in particular for the questions that are more open-ended.
2 Problems

1. Suppose that the integers from 1 to \( n \) are permuted (uniformly at random) and then inserted one at a time to construct a standard binary search tree. (If you don’t know what a standard binary search tree is, you may skip this problem or feel free to look it up; it only affects the first part of the problem.) Call this experiment A. We would like to show that in this case the depth of the tree is \( O(\log n) \) with high probability. We show how to do this via an interesting exponential embedding.

(a) Think of a binary tree growing in the following manner. Start with a root (at depth 0) at time 0. Each node can give birth to two children (the left and the right children). Once a node is born, the time until that node births each of its children is an independent, exponentially distributed random variable with mean 1. We stop once \( n \) nodes are in the tree. Explain why the distribution of the shape of the resulting binary tree is equivalent to the distribution on the shape of the binary tree constructed in experiment A.

(b) Show that the expected time until there are \( n \) nodes in the tree is \( H(n) - 1 \).

(c) Show that with probability at least \( 1 - 1/n \), there are \( n \) nodes in the tree before time \( 10 \log_2 n \). (Hint, use some type of Chernoff bound.)

(d) Let \( X_1, X_2, \ldots \) be independent exponentially distributed random variables with mean 1. Show that

\[
\Pr( \sum_{i=1}^{c_1 \log_2 n} X_i \leq 10 \log_2 n) \leq 1/n^{c_1+1}
\]

for some \( c_1 \geq 1 \).

(e) Now argue that the probability that the depth of the tree is at least \( c_1 \log_2 n \) is at most \( 2/n \).

2. Suppose that \( n \) married couples are seated at a round table at a dinner party. The \( 2n \) people are seated by giving each person a seat chosen uniformly at random.

(a) Find the expected number of couples where the partners are seated next to each other as a function of \( n \).

(b) Find the variance of the number of couples where the partners are seated next to each other as a function of \( n \).

(c) Show that the probability that the number of couples where the partners are seated next to each other is 0 is strictly less than \( 1/2 \).

3. A \( k \)-uniform hypergraph is an ordered pair \( G = (V, E) \), but edges consist of sets of \( k \) (distinct) vertices, instead of just 2. (So a 2-uniform hypergraph is just what we normally call a graph.) A hypergraph is \( k \)-regular if all vertices have degree \( k \); that is, they are in \( k \) hypergraph edges.

Show that for sufficiently large \( k \), the vertices of a \( k \)-uniform, \( k \)-regular hypergraph can be 2-colored so that no edge is monochromatic. What’s the smallest value of \( k \) you can achieve? (Hint: Think LLL.)

4. We throw \( kn \) red balls and \( kn \) blue balls uniformly at random into \( n \) bins. Here \( k > 1 \) is a constant, independent of \( n \). A bin is balanced if the number of red balls and blue balls that land in the bin is equal.

(a) Find the expected number of balanced bins as a function of \( n \) and \( k \). (Note: I do not expect a pretty closed-form formula for this. If you happen to find one, all the better.)

(b) Show that (for all sufficiently large \( n \) and \( k \)) the number of balanced bins is at least \( cn/\sqrt{k} \) for some constant \( c \) independent of \( n \) and \( k \) with probability at least \( 1 - 1/n^3 \).

5. Suppose we are given \( n \) records, \( R_1, R_2, \ldots, R_n \). The records are kept in some order. The cost of accessing the \( j \)th record in the order is \( j \). So if we had 4 records, and they were ordered as \( R_2, R_4, R_3, R_1 \), the cost of accessing \( R_4 \) would be 2, the cost of accessing \( R_1 \) would be 4, etc.

Suppose that at each step, record \( R_j \) is accessed with probability \( p_j \), with each step being independent of other steps. If we knew the values of the \( p_j \) in advance, we would keep the \( R_j \) in decreasing order with respect to \( p_j \). But if we don’t know the \( p_j \) in advance, we might use the move-to-front heuristic: at each step, put the record that was accessed at the front of the list. We assume this can be done with no cost, and that all other records stay in the same order. For example, if the order was \( R_2, R_4, R_3, R_1 \) and \( R_3 \) was accessed, at the next step the order would be \( R_3, R_2, R_4, R_1 \). In this setting, the order of the records can be thought of as the state of a Markov chain.
(a) Find the stationary distribution of this chain. (Explain for an ordering of records how the stationary distribution could be computed.)

(b) Let $X_k$ be the cost for accessing the $k$th requested record. Determine an expression for $\lim_{k \to \infty} E[X_k]$. Your expression should be computable in time polynomial in $n$ given the $p_j$’s.

6. Consider the numbers modulo $n$. We describe a Markov chain on the numbers as follows: at each step, we add 1 (with probability 1/2) or subtract one (with probability 1/2) from the current number. Suppose we start at 0, and ask what is the last number that is reached by this walk. For example, working modulo 5, if the sequence of states is 0 1 0 4 3 4 3 2 ..., then 2 is the last number reached. Notice that 0 can never be the last number reached since we begin there. Given $n$, what is the most likely number to be last reached? Prove your answer.

7. Consider all $\binom{n}{k}$ subsets of size $k$ of $\{1, \ldots, n\}$, $T_1, T_2, \ldots, T_{\binom{n}{k}}$. We want to find the smallest collection of subsets $S_1, S_2, \ldots, S_m$ of $\{1, \ldots, n\}$, each of size exactly $\ell > k$, so that each $T_i$ is a subset of at least one of the $S_j$. When a $T_i$ is the subset of some $S_j$ we say that it is covered.

   (a) Argue that $m \geq \binom{n}{k}/\binom{\ell}{k}$.

   (b) Let $m_0 = \binom{n}{k}/\binom{\ell}{k}$. Suppose we choose $m_1 = m_0 \ln(\binom{\ell}{k})$ subsets of $\{1, \ldots, n\}$ of size $\ell$ independently and uniformly at random. Show that the expected number of $T_i$ that are not covered is at most $m_0$.

   (c) Argue that there exists a solution with $m \leq m_0(1 + \ln(\binom{\ell}{k}))$. 