A *practice* exam starts on the next page. We've done our best to draw this exam independently from the same distribution[1] as the real exam. Inspired by this, here is a warm-up question about how best to use this practice exam.

(Note: this is not a serious question :))

0. **(0 pt.)** Suppose that a class covers n topics. Each student i in the class has studied up on a set $S(i) \subseteq \{1, \ldots, n\}$ of topics. To create the exam, the instructors choose k topics to test, independently and uniformly at random (with replacement) among the n topics, and give $100/k$ points to each. Suppose that n and k are sufficiently large[2], and that $k = o(n)$.

   (a) What is the expected score of student i, in terms of $|S(i)|$? (Assuming that student i aces any problem on a topic they have studied).

   (b) Suppose that a student wants to use the practice exam to assess if they have studied enough. That is, the student studies first, and then scores s on the practice exam taken under the same conditions as the actual final. Bound the probability that they score worse than $s - 1$ (out of 100) on the actual final, asymptotically in terms of k and/or n.

   (c) Suppose on the other hand that a student wants to use the practice exam to help guide their study. That is, they student studies the topics that appear on the practice exam as they take it. Say the student scores s on the practice exam taken in this open-book way. What can you say about how well the student will score on the actual final?

   (d) How many topics should a student study up on to maximize their expected score on the final exam?

SOLUTION:

(a) $\mathbb{E}[\text{score of student } i] = \sum_{j=1}^{k} \frac{100}{k} \mathbb{P}\{j^{th} \text{ exam topic} \in S(i)\} = 100 \cdot \frac{|S(i)|}{n} =: \mu_i$

(b) For both the real and the practice exam,

$$\mathbb{P}[|\text{score}_i - \mu_i| > t] \underset{\text{Hoeffding}}{\leq} 2\exp\left(-t^2 \Big/ 2\sum_{j=1}^{k} \left(\tfrac{100}{k}\right)^2\right) = 2\exp\left(\frac{-t^2 k}{20000}\right)$$

If we choose $t = 1/2$ and k is sufficiently large, then whp $|\text{practice score}_i - \text{real score}_i| \leq \frac{1}{2} + \frac{1}{2} = 1$.

(c) You can't say much, since the exams are independent. The set of topics could be totally disjoint!

(d) All of them!

[1] Okay, to be honest, this practice exam is not as vetted as a real exam would be, and it might involve a bit more reading (i.e., the problems take longer to state) than we'd ideally put on a timed exam. So it's not quite the same distribution. But we are shooting for the same distribution of, say, difficulty.

[2] It seems reasonable to us that there would be upwards of a million topics on the exam...

Instructions

- **DO NOT OPEN THE EXAM UNTIL YOU ARE INSTRUCTED TO.** *(Note: since this is a practice exam, go ahead when you feel ready!)*
- Answer all of the questions as well as you can. You have three hours to complete this exam.
- The exam is **non-collaborative**; you must complete it on your own. If you have any clarification questions, please ask the course staff (we are outside the exam room). We cannot provide any hints or help.
- This exam is **closed-book**, except for:
  - **Up to three double-sided sheets of paper** that you have prepared ahead of time. You can have anything you want written on these sheets of paper.
  - We have also provided a "cheat-sheet" with some helpful theorems and inequalities. **You can find this as the last page of this exam. Feel free to rip it off of the exam.**
- **Please DO NOT separate pages of your exam** (except for the cheat sheet at the back). The course staff is not responsible for finding lost pages, and you may not get credit for a problem if it goes missing.
- There are a few pages of extra paper at the back of the exam in case you run out of room on any problem. If you use them, please clearly indicate on the relevant problem page that you have used them, and please clearly label any work on the extra pages.

General Advice

- If you get stuck on a question or a part, move on and come back to it later. The questions on this exam have a wide range of difficulty, and you can do well on the exam even if you don't get a few questions.
- Pay attention to the point values. Don't spend too much time on questions that are not worth a lot of points.
- There are 100 points total on this exam.

Name (please print clearly):

SOLUTIONS
____

Honor Code

The following is a statement of the Stanford University Honor Code:

1. *The Honor Code is an undertaking of the students, individually and collectively:*
   (1) that they will not give or receive aid in examinations; that they will not give or receive unpermitted aid in class work, in the preparation of reports, or in any other work that is to be used by the instructor as the basis of grading;
   (2) that they will do their share and take an active part in seeing to it that others as well as themselves uphold the spirit and letter of the Honor Code.
2. *The faculty on its part manifests its confidence in the honor of its students by refraining from proctoring examinations and from taking unusual and unreasonable precautions to prevent the forms of dishonesty mentioned above. The faculty will also avoid, as far as practicable, academic procedures that create temptations to violate the Honor Code.*
3. *While the faculty alone has the right and obligation to set academic requirements, the students and faculty will work together to establish optimal conditions for honorable academic work.*

For my part, I believe that we have upheld our end of the agreement in Item 2. I don't think we are taking unusual or unreasonable precautions that would indicate a lack of confidence in the honor of students, and I believe that the in-person setting avoids temptations ot violate the honor code to the extent practicable.

[signed, Mary Wootters]

Please acknowledge that you have held up your end of the agreement in Item 1:

I have abided by the Honor Code, and in particular the policies listed above, both in letter and in spirit, while taking this exam.

signed, ________________________________

Good Luck!

Note: (b) and (c) on this question could take a long time if you write out all the details. On a real timed exam we would not ask you to do that.

1. **(26 pt.)**

(a) **(7 pt.)** Let $\varepsilon \in (0, 1/2)$. Suppose you have an algorithm A that outputs answers in $\{0,1\}$ and is correct with probability $\frac{1}{2}+\varepsilon$. You decide to make a more robust algorithm, $\tilde{A}$, that just runs A independently T times and returns the most frequent answer. Show that $\tilde{A}$ is correct with probability 0.99 for some value T that is $O(1/\varepsilon^2)$.

Let $X_i = \mathbb{1}\{i^{th} \text{ run of } A \text{ is correct}\}$.

By Chebyshev's inequality, $\mathbb{P}\left[\left|\sum_{i=1}^{T} X_i - T(\tfrac{1}{2}+\varepsilon)\right| > \varepsilon T\right] \leq \frac{T(\frac{1}{2}+\varepsilon)(\frac{1}{2}-\varepsilon)}{\varepsilon^2 T^2} \leq \frac{1}{4\varepsilon^2 T}$.

Thus, if $T \geq 25/\varepsilon^2$, this 'probability' is ≤ 0.01, and a majority of the trials are correct. [Note: you could also use a Chernoff bd here].

(b) **(7 pt.)** Now suppose that A can output answers in $\{0, 1, \ldots, n\}$, instead of just $\{0,1\}$. Suppose that A is correct with probability at least $p \geq \frac{C \log n}{n \log\log n}$, where C is some constant that you get to choose. Further suppose that for any *incorrect* answer $i \in \{0, 1, \ldots, n\}$, the probability that A outputs i is at most $1/n$. ~~Show~~ SKETCH a proof that, for sufficiently large n, $\tilde{A}$ (which still returns the most frequent answer out of the T trials) is correct with probability at least 0.99 when $T = n$.

First, I'd use a Chernoff bound to show that with probability $\geq 1-o(1)$, at least $pn/2 = \frac{C\log(n)}{2\log\log(n)}$ of the trials output the correct answer.

Second, for any one incorrect answer, i, the probability that i gets k votes is at most

$$\underbrace{\binom{n}{k}}_{\text{union bound over all possible } k \text{ trials that could vote for } i} \cdot \underbrace{\left(\frac{1}{n}\right)^k}_{\text{bound on the probability that all } k \text{ of those vote for } i} \leq \frac{n^k}{k!}\left(\frac{1}{n}\right)^k = \frac{1}{k!} \leq \left(\frac{e}{k}\right)^k$$

Since we asked for a sketch, you could be a bit more handwavey than this. Note that this is basically the same arg. we used to show that the max load of n balls into n bins is $O\left(\frac{\lg n}{\lg\lg n}\right)$ whp, so you could have just appealed to that.

For any $k \geq \frac{C\log(n)}{2\log\log n}$, this is $\leq \left(\frac{e}{k}\right)^k \leq \left(\frac{2\log\log(n)}{C\log n}\right)^{\frac{C\log(n)}{2\log\log n}} = \exp\left[\frac{-C\log(n)}{\log\log(n)}\left[\log\log(n) + \lg\left(\frac{C}{2\lg\lg n}\right)\right]\right]$

$= \exp\left(-C\log(n)\left[1+o(1)\right]\right)$

$\approx (1/n)^C$

Then we union bound over all $k = \frac{pn}{2}, \ldots, n$ and over all n values of i (our incorrect answer) and conclude

$$\mathbb{P}\left[\text{any incorrect answer gets} > \tfrac{pn}{2} \text{ votes}\right] \leq n^2 \cdot \left(\frac{1}{n}\right)^C,$$

So if we choose $C \geq 3$ this is [easily] $o(1)$.

[More parts on next page]

Then we union bound over the "First" and "Second" paragraphs to conclude that with prob. $1-o(1)$, the right answer gets $\geq \frac{np}{2}$ votes, while all the incorrect answers get $< \frac{np}{2}$ votes.

Whoops! There's a typo here, it should be Poi(1/2), and then all the 2's should be replaced with 1/2's going forward. (The proof stays the same though).

[Continued from previous page]

(c) **(7 pt.) Sketch a proof** that shows that, in the previous part, you cannot take p to be substantially smaller. That is, explain why there are some constants $C', n_0 > 0$ so that, for any $n \geq n_0$, if $p \leq \frac{\log n}{C' n \log\log n}$, then when $T = n$, $\tilde{A}$ could be incorrect with probability at least $1/2$.

Note: you don't need to give a super formal proof, but explain the steps you would go through to give a formal proof. (That is, what theorems/inequalities would you use, on what random variables, and how would they fit together to prove this result?)

Again, this is basically the same arg. showing that max load is $\Omega\left(\frac{\log n}{\lg\lg n}\right)$ whp. As before, you didn't need to go into this much detail, you could have appealed to that argument.

Suppose that "0" is the correct answer (prob. p), and that the incorrect answers $1, \ldots, n$ are all equally likely (prob $\frac{1-p}{n}$).
By a Chernoff bd, w/ prob $1 - o(1)$ there are $<$ $2pn$ votes for the correct answer.
Conditioned on that, we can view the remaining $\geq (1-2p)n$ votes as "balls" into our n wrong answers ("bins").
We approximate this w/ Poissonization: imagine dropping $k \sim \mathrm{Poi}\left(\frac{n}{2}\right)$ balls instead. With prob. $1 - o(1)$, by our Poisson tail bound, $k \leq n(1-2p)$, so if there are $> 2pn$ votes for any incorrect answer in the Poissonized setting, there will be in our setting as well. Then the # votes for incorrect answer i is a Poisson random variable, $X_i \sim \mathrm{Poi}(2)$, and the X_i are independent. At this point we can use the def. of Poisson random vars to show that

$$\mathbb{P}[X_i \geq 2pn] \geq \mathbb{P}[X_i = 2pn] = \frac{e^{-2} 2^{2pn}}{(2pn)!} \geq \frac{1}{e^2} \cdot 2^{2pn} \left(\frac{1}{2pn}\right)^{2pn} = \exp(2 - 2pn \cdot \log(pn))$$
$$= \exp\left(2 - 2\frac{\log n}{C' \log\lg(n)} \cdot O(\lg\lg n)\right)$$
$$= \exp\left(2 - \frac{2C'' \log(n)}{C'}\right) \text{ for some const. } C''.$$

For large enough C', this is, say, $\geq 1/\sqrt{n}$.
By independence of the X_i's, $\mathbb{P}[\exists i, X_i \geq 2pn] \geq 1 - (1 - 1/\sqrt{n})^n \approx 1 - \exp(-\sqrt{n}) \geq 1/2$, as desired.

(d) **[May be more difficult] (5 pt.)** As above, say that A outputs answers in $\{0, 1, \ldots, n\}$. Now suppose that A is correct with probability $1/4$, and can output any particular incorrect answer i with probability at most $1/8$. How small can you take T to still allow the guarantee that $\tilde{A}$ is correct with probability at least 0.99?

T could be as small as $O(1)$.
Say WLOG that the "correct" answer is 0. By a Chernoff bound,

$$\mathbb{P}[\#\text{votes for } 0 < T/5] \leq \exp\left(\frac{-(\frac14 - \frac15)^2 T/4}{3}\right) = \exp\left(-T/C\right) \text{ for some constant } C. \quad (*)$$

Break up the incorrect answers $1, \ldots, n$ into chunks $S_1, \ldots, S_r \subseteq \{1, \ldots, n\}$ ($S_i \cap S_j = \emptyset$) so that $\forall j$, $\sum_{i \in S_j} \mathbb{P}[A \text{ outputs } i] \in [1/16, 3/16]$. We can do this via the greedy algorithm: Keep taking values i into S_j until the first time $\sum_{i \in S_j} \mathbb{P}[A \text{ says } i] > 1/16$. Since each of these probabilities is $\leq 1/8$, we'll have $\sum_{i \in S_j} \mathbb{P}[A \text{ says } i] \leq 1/16 + 1/8 = 3/16$.
Note that $r \leq 16$, or else $\sum_{i=1}^n \mathbb{P}[A \text{ outputs } i] > 1$.
Fix some chunk $S = S_j$. By a Chernoff bound, (say, Bernstein's ineq.)

$$\mathbb{P}\Big[\underbrace{\#\text{votes for any } i \in S}_{\substack{\text{this is the sum of } T \text{ indep. r.v.'s,} \\ \text{and } \mathbb{E} \text{ is } \leq 3/16\, T}} \geq \left(\frac{3}{16} + \frac{1}{32}\right)T\Big] \leq \exp\left(\frac{-(T/32)^2}{T}\right) = \exp\left(-T/32^2\right). \quad (**)$$

If we choose T to be a big enough constant, both $(*)$ and $(**)$ are $\leq \frac{0.01}{17}$, and then by a union bound over $(*)$ and all $r \leq 16$ instantiations of $(**)$ we see that there are more votes for 0 than any other answer w/ prob ≥ 0.99.

2. **(24 pt.)** For each of the following tasks, **briefly** sketch a randomized algorithm that does it and **briefly** explain why it works. You do not need to give a formal proof that it works. You can use any algorithm we have seen in class as a black box (unless otherwise noted), and your answer should be no more than a few sentences and possibly some very short pseudocode for each part.

   (a) **(6 pt.)** Given a connected, undirected, unweighted graph G on n vertices and m edges, find a cut $(S, \bar{S})$ so that the number of edges crossing the cut is minimized, with probability at least 0.9. The algorithm should run in time poly(n).

   Use Karger's algorithm.

   (b) **(6 pt.)** Given a data set $X \subseteq \mathbb{R}^N$ of size N with $\|x\|_2 = 1$ for all $x \in X$, give a randomized algorithm that returns estimates of $\|x - y\|_2$ for *all* pairs $x, y \in X$. With probability at least 0.99, your estimates should all be accurate up to a multiplicative factor of (1 ± 0.01). Your algorithm should run in time $O(N^2 \log N)$.

Let $A \in \mathbb{R}^{m \times N}$ be a matrix of iid Gaussians for $m = O(\log N)$, as guaranteed by JL lemma w/ $\varepsilon = 0.01$.

For all $x \in X$, compute $A \cdot x$ // takes time $O(\overset{N \text{ pts}}{N} \cdot \underbrace{N \log N}_{\text{time for matrix-vector mult.}}) = O(N^2 \log N)$

For all $x, y \in X$, compute $\|Ax - Ay\|_2$ // time $O(N^2 \cdot \log N)$

$\binom{N}{2}$ pairs — takes time $\log N$ to compute distances btwn vectors of length $\log N$.

by JL lemma, $\|Ax - Ay\|_2 = (1 \pm 0.01)\|x - y\|_2$

Return these as estimates.

[More parts on next page]

[Continued from previous page]

(c) **(6 pt.)** Say you are given a 2-CNF formula φ (that is, φ is of the form $(x_1 \vee x_2) \wedge (\overline{x_1} \vee x_3) \wedge \cdots$, with n variables and m clauses that each contain two distinct literals). Describe a randomized algorithm that runs in time $\text{poly}(n, m)$ and, if there is a satisfying assignment to φ, returns it with probability at least 0.99. If there is no satisfying assignment, your algorithm should return "NOPE" with probability 1.

Note: Please describe your algorithm, don't use an algorithm from class as a black box.

This is the 2SAT problem, and we saw an algorithm to solve it in class:

Start w/ a random assignment σ

For $t = 1, \ldots, 100\, n^2$:

- If there is some clause involving x_i, x_j that is not satisfied, choose $r \in \{i, j\}$ uniformly at random and flip x_r in σ
- Else return σ

Return NOPE

(d) **(6 pt.)** Let (X, d) be an arbitrary finite metric space with $|X| = n$. We say that k points $x_1, \ldots, x_k \in X$ form an *r-cluster* if $d(x_i, x_j) \leq r$ for all $i, j \in \{1, \ldots, k\}$. For a general metric space (X, d), the problem of finding an r-cluster of size k, given k and r, seems pretty hard; but fortunately you have access to a magic genie who can do it for $(\mathbb{R}^d, \ell_1)$ in time polynomial in n and in $2^{\sqrt{d}}$. (If there is no r-cluster of size k, the genie outputs "Sorry, no such cluster.")

Give a randomized algorithm that runs in time $\text{poly}(n)$, takes as input k and r, and satisfies the following guarantee with probability at least 0.99, for some $D = O(\log n)$: If there is r/D-cluster in X of size k, the algorithm must output an r-cluster in X of size k. (Otherwise it can do whatever it wants).

Use Bourgain's embedding to embed X into $(\mathbb{R}^{O(\log^2 n)}, \ell_1)$ with distortion $O(\log(n))$.

Then use the genie to find clusters in the embedding, and return those.

3. **(15 pt.)** Let G be a cycle with $\cancel{11}$ 16 n vertices. (That is, the vertices are labeled $0, 1, \ldots, \cancel{11}$ 16 $n-1$, and each vertex i is connected only to $i \pm 1 \bmod \cancel{11}$ 16 n). The vertices are each colored one of n colors, with $\cancel{11}$ 16 occurrences of each color. Show that it is always possible to find n vertices in G so that all are distinct colors, and no two are connected by an edge.

We'll use the LLL!

For each color $i \in [n]$, choose a random vertex v_i of color i.

For each edge $e \in E(G)$, let A_e be the bad event that both endpoints of e were selected.

Then $\mathbb{P}[A_e] = \begin{cases} 0 & \text{endpts of } e \text{ had the same color} \\ \left(\frac{1}{16}\right)^2 & \text{otherwise} \end{cases} \leq \frac{1}{16^2}$

Further, A_e is mutually independent of $A_{e'}$ for all e' whose endpoints don't share a color with either endpoint of e.

The number of edges w/ endpts that DO share a color with either endpt of e is at most 62. That's because for each of the ≤ 2 colors, there are 16 vertices of that color, and each of them have ≤ 2 edges touching them. That's 64 edges, but we've counted e itself twice so we can subtract 2 to get 62. Thus, we can take $d = 62$.

Now, we have

$$p \cdot d = \frac{62}{16^2} < \frac{64}{16^2} = \frac{4}{16} = \frac{1}{4}.$$

So we can apply the LLL and conclude that there's a way to choose vertices so that none of the A_e occur.

4. **(15 pt.)** Consider the following procedure for shuffling a deck of n cards: Choose two indices $i, j \in \{1, \ldots, n\}$ uniformly and independently at random, and switch the card at position i and the card at position j. (Note that it is possible that $i = j$).

In this problem, we will use a coupling argument to bound the mixing time τ_{mix} of this procedure. Let X_t denote the state of the deck after we have swapped t cards.

(a) **(5 pt.)** Your friend suggests the following coupling (which we also encountered on a quiz). Let X_t be the walk described above, and define Y_t to be a walk that makes the *same* choice of i and j at each step. Unfortunately, this isn't a great idea for bounding τ_{mix}. In at most a few sentences, explain why not.

This walk will never couple, unless $X_0 = Y_0$.

(b) **(10 pt.)** Here's another coupling to consider. We will view our shuffling procedure slightly differently: Instead of choosing i, j at random, choose a *card* c (like "the ace of spades") uniformly at random, and choose an index $i \in \{1, \ldots, n\}$ uniformly at random. Then switch the card c with whatever card at index i. Note that this is an alternative way of defining the same Markov chain $\{X_t\}$. Now define a coupling (X_t, Y_t) by choosing the *same* choice of c and i in both chains.

Use this coupling to show that $\tau_{mix} = O(n^2)$.

[**HINT:** *Keep track of a variable D_t which is defined to be the number of positions in which the decks X_t and Y_t differ. Show that* $\Pr[D_{t+1} < D_t] \geq (D_t/n)^2$.]

Let D_t be as in the hint.

Consider the following cases:

CASE 1. Card c is already in the same place in decks 1 and 2. ← Probability $1 - \frac{D_t}{n}$, because we had to pick c to be one of the $n - D_t$ cards where the 2 decks agree.

In this case, $D_{t+1} = D_t$.

CASE 2. Card c is in different places in the two decks. ← Prob. $\frac{D_t}{n}$

Case 2a The card at location i is the same in both decks ← Prob $\frac{D_t}{n} \cdot \left(1 - \frac{D_t}{n}\right)$

In this case, $D_{t+1} = D_t$ (we aligned the "c"'s, but messed up what was @ index i)

Case 2b The card at location i is different between the 2 decks. ← Prob. $\frac{D_t}{n} \cdot \frac{D_t}{n}$

In this case, $D_{t+1} = D_t - 1$ (we aligned the "c"'s, and didn't mess up anything else).

[More space on next page]

continued

[Continued from previous page; more space for part (b)]

By the analysis above,

$$\mathbb{P}[D_{t+1} = D_t - 1] = \left(\frac{D_t}{n}\right)^2, \quad \text{and } \mathbb{P}[D_{t+1} = D_t] = 1 - \left(\frac{D_t}{n}\right)^2.$$

Thus, the expected amount of time Δ to go from $D_t = i$ to $D_{t+\Delta} = i-1$ is $\left(\frac{n}{i}\right)^2$.

Let T be the time needed to go from $D_0 \le n$ to $D_T = 0$. By linearity of $\mathbb{E}$,

$$\mathbb{E}T = \left(\frac{n}{n}\right)^2 + \left(\frac{n}{n-1}\right)^2 + \left(\frac{n}{n-2}\right)^2 + \cdots + 1^2 = n^2 \sum_{j=1}^{n} \frac{1}{j^2}$$

$$\le n^2 \underbrace{\sum_{j=1}^{\infty} 1/j^2}_{\text{this sum converges.}} = O(n^2)$$

Then by Markov's inequality,

$$\mathbb{P}[T > 2e(\mathbb{E}T)] \le 1/2e.$$

Then $\Delta(2e \cdot \mathbb{E}[T]) \le \mathbb{P}[T > 2e\,\mathbb{E}T] \le 1/2e$,

so $\tau_{mix} \le 2e \cdot \mathbb{E}[T] = O(n^2)$, as desired.

(c) **(0 pt.) BONUS** [We wouldn't put this on a real exam, but it might be fun to think about :)]. Show that $\tau_{mix} = O(n \log n)$.

Check out, e.g., "A strong stationary time for random transpositions" [Graham White, 2019] and the references therein.

5. **(20 pt.)** Let $s_1, s_2 \in \{0,1\}^n$ denote two independent and uniformly random length n Boolean strings.

A *subsequence* of a string $s \in \{0,1\}^n$ is any sequence of the form $s[i_1]s[i_2]\cdots s[i_\ell]$ for $i_1 < i_2 < \cdots < i_\ell$. For example, 000 is a subsequence of 010101. A *common subsequence* between strings s_1 and s_2 is a subsequence that's common to both. For example, $s_1 = 010101$ and $s_2 = 001100$ have a common subsequence 000. They also have a longer common subsequence, 0011. The *longest common subsequence* is a common subsequence with the most bits in it: in this example, one happens to be 0011.

(a) **(10 pt.)** Letting L denote the length of the longest common subsequence of the two strings, prove that $\Pr[|L - \mathbb{E}[L]| \geq \lambda] \leq 2e^{-\frac{\lambda^2}{2n}}$.

We'll use Azuma-Hoeffding. Let $Z_t = \mathbb{E}[L \mid X_0, \ldots, X_t]$ where $X_t \in \{0,1\}^2$ is the $X_t = (s_1[t], s_2[t])$. be the Doob martingale. Then

$|Z_t - Z_{t-1}| \leq 1$, because one additional bit can change the length of the LCS by ≤ 1.

So by Azuma,

$$\mathbb{P}[|Z_n - Z_0| \geq \lambda] \leq 2\exp\left(\frac{-\lambda^2}{2n}\right).$$

this is $\mathbb{E}[L]$

this is L

(b) **(5 pt.)** Prove that for sufficiently large n, with probability tending to 1 as $n \to \infty$, the length of the longest common subsequence of the two strings is at least $0.49n$.

We will show that w/ prob $1-o(1)$, both strings have at least $0.49n$ zeros.

Indeed, let $X_i = 1$ iff $s_1[i] = 0$, so the X_i are iid Ber(1/2).

Then by a Chernoff bound,

$$\mathbb{P}\left[\sum_i X_i < 0.49n\right] \leq \exp\left(\frac{-(0.01)^2 \cdot 0.49n}{3}\right)$$

$$= \exp(-\Omega(n)).$$

The same holds for s_2.

By a union bound, whp both s_1 and s_2 have $\geq 0.49n$ zeros, so the LCS is at least that long.

(Indeed, $\underbrace{000 \cdots 0}_{0.49n}$ is a common subsequence).

[Another part on next page]

Note: Pinning down $\mathbb{E}[L]$ exactly is open! See "Improved Bounds on the avg length of LCS's" [Lueker 2009] and the references therein for an overview of what's known.

[Continued from previous page]

(c) **(5 pt.)** **[May be more difficult]** Find a constant $c > 1/2$ and prove $\mathbb{E}[L] \geq cn$.

There are lots of ways to do this. One way is to observe that for $n=2$, $\mathbb{E}[L] > \frac{1}{2}n$, and then break up our string into chunks of size 2 and stitch together the LCS's from each chunk. To see that this holds for $n=2$, we can just count:

S2 \ S1	00	01	10	11
00	2	1	1	0
01	1	2	1	1
10	1	1	2	1
11	0	1	1	2

eg, this is LCS(01, 11)

Thus, $\mathbb{E}[L] = \dfrac{4 \cdot 2 + 10 \cdot 1}{16} = \dfrac{18}{16} = 9/8$

So we can take our constant c to be at least $\dfrac{9}{2 \cdot 8} = \dfrac{9}{16} > \dfrac{1}{2}$.

This is the end!

This is the end of the exam! You can use this page for extra work on any problem. **Keep this page attached** to the exam packet, and if you want work on it graded, clearly label which question your extra work is for.

This page is for extra work on any problem. **Keep this page attached** to the exam packet, and if you want work on it graded, clearly label which question your extra work is for.

This page is for extra work on any problem. **Keep this page attached** to the exam packet, and if you want work on it graded, clearly label which question your extra work is for.

Some useful inequalities, definitions and theorem statements

Note: We have not always stated full theorems here, just the quantitative punchlines. You are responsible for knowing when each theorem applies.

Inequalities and Series

- $1 - x \leq e^{-x}$ for any x.
- $(n/k)^k \leq \binom{n}{k} \leq (en/k)^k$ for all $k \leq n$.
- $\binom{n}{k} \leq \frac{n^k}{k!}$ for all $k \leq n$.
- $\sum_{i=1}^{n} 1/i = \Theta(\log n)$
- $\sum_{i=1}^{n} 1/i^c = O(1)$ for all $c > 1$.

Definitions

- $f(n) = O(g(n))$ means that there are some constants $c, n_0 > 0$ so that for all $n \geq n_0$, $f(n) \leq cg(n)$.
- $f(n) = \Omega(g(n))$ means that there are some constants $c, n_0 > 0$ so that for all $n \geq n_0$, $f(n) \geq cg(n)$.
- $f(n) = o(g(n))$ means that $\frac{f(n)}{g(n)} \to 0$ as $n \to \infty$.
- $f(n) = \omega(g(n))$ means that $\frac{f(n)}{g(n)} \to \infty$ as $n \to \infty$.
- If $X \sim \mathrm{Poi}(\lambda)$, then $\Pr[X = k] = \frac{e^{-\lambda}\lambda^k}{k!}$.
- If $X \sim N(\mu, \sigma^2)$, then $\Pr[X = x] = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$
- If $X \sim \mathrm{Ber}(p)$, then $X \in \{0, 1\}$ and $\Pr[X = 1] = p$.

Concentration Inequalities

- Markov's inequality: For a non-negative random variable X, $\Pr[X > t] \leq \frac{\mathbb{E}X}{t}$.
- Chebyshev's inequality: For any random variable X, $\Pr[|X - \mathbb{E}X| > t] \leq \frac{\mathrm{Var}(X)}{t^2}$.
- A few Chernoff bounds: For independent $X_i \in \{0, 1\}$, if $X = \sum_{i=1}^{n} X_i$, then:
  - For $\delta > 0$, $\Pr[X \geq (1 + \delta)\mathbb{E}[X]] \leq \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^{\mathbb{E}[X]}$. If $\delta \in (0, 1]$ this is $\leq \exp(-\delta^2\mathbb{E}[X]/3)$.
  - For $\delta \in (0, 1]$, $\Pr[X \leq (1-\delta)\mathbb{E}[X]] \leq \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mathbb{E}[X]}$. If $\delta \in (0, 1]$, this is $\leq \exp(-\delta^2\mathbb{E}[X]/2)$.
  - For $c \geq 6$, $\Pr[X \geq c\mu] \leq 2^{-c\mu}$.

- Tail bound for Poisson random variables: If $X \sim \text{Poi}(\lambda)$, then for any $c > 0$, $\Pr[|X - \lambda| \geq c] \leq 2\exp\left(\frac{-c^2}{2(c+\lambda)}\right)$.

- Azuma-Hoeffding Inequality: Let $\{Z_t\}$ be a martingale with respect to $\{X_t\}$, and suppose $|Z_i - Z_{i-1}| \leq c_i$ for all $i \leq n$. For any $\lambda > 0$, $\Pr[|Z_n - Z_0| \geq \lambda] \leq 2\exp\left(\frac{-\lambda^2}{2\sum_{i=1}^n c_i^2}\right)$.

Dimension Reduction

- Bourgain's Embedding: for any finite metric space (X, d) with $|X| = n$, there is an embedding of (X, d) into $\mathbb{R}^k$ under the ℓ_1 metric with distortion $O(\log n)$, where $k = O((\log n)^2)$.

- Johnson-Lindenstrauss Lemma: for any $\varepsilon \in (0, 1)$, for any $X \subseteq \mathbb{R}^d$ with $|X| = n$, there is a linear map $f : \mathbb{R}^d \to \mathbb{R}^m$ with $m = O(\varepsilon^{-2} \log n)$ that embeds (X, ℓ_2) into $(\mathbb{R}^m, \ell_2)$ with distortion at most $(1 + \varepsilon)$.

Probabilistic Method

- Second moment method: for real-valued X, $\Pr[X = 0] \leq \frac{\text{Var}[X]}{(\mathbb{E}[X])^2}$.

- LLL: Let $A_1, \ldots, A_n$ be events so that $\Pr[A_i] \leq p$ for all i, and where each A_i is mutually independent of all but d other events. Then:

  - If $pd \leq 1/4$, then $\Pr[\bigcap_i \overline{A_i}] > 0$
  - If $p(d+1) \leq 1/e$, then $\Pr[\bigcap_i \overline{A_i}] > 0$.

Markov Chain / Martingale Theorems

- Fundamental theorem of Markov chains: Let $\{X_t\}$ be an irreducible aperiodic Markov chain over a finite state space with transition matrix P. Then there is a unique stationary distribution π so that $\Pr[X_t = i | X_0 = j] \to \pi_i$ for all states i, j. Further, π_i is the expected return time of state i, and $\pi P = \pi$.

- Let $\{X_t\}$ be a finite irreducible aperiodic Markov chain with a coupling $\{(X_t, Y_t)\}$. Then $\Delta(t) \leq \max_{s,s'} \Pr[X_t \neq Y_t | X_0 = s, Y_0 = s']$.

- Let $\{X_t\}$ be a finite irreducible aperiodic Markov chain and let T be a strong stationary stopping time. Then $\Delta(t) \leq \Pr[T > t]$.

- The *Doob Martingale* for a quantity A is $Z_t = \mathbb{E}[A | X_0, \ldots, X_t]$. Theorem: it is a martingale.

- Martingale stopping theorem: Let $\{Z_t\}$ be a martingale with respct to $\{X_t\}$. Let T be a stopping time for $\{X_t\}$. Then $\mathbb{E}[Z_T] = \mathbb{E}[Z_0]$ if at least one of the following holds:

  1. There is a constant c s.t. $|Z_i| \leq c$ for all i.
  2. There is a constant c s.t. $T < c$ with probability 1.
  3. $\mathbb{E}[T] < \infty$ and there is a constant c s.t. for all i, $\mathbb{E}[|Z_{i+1} - Z_i| | X_0, \ldots, X_i] < c$.