CS 574, Spring 2021
Timothy Chan

Homework 3 (due March 31 Wednesday 5pm (CT))

Instructions: You may work in groups of at most 3; submit one set of solutions per group. Always acknowledge any discussions you have with other people and any sources you have used (although most homework problems should be doable without using outside sources). In any case, *solutions must be written entirely in your own words.*

1. [*28 pts*] We are given a text string $A[1]A[2]\cdots A[n] \in \{0,1\}^*$ of length n and a pattern string $B[1]B[2]\cdots B[m] \in \{0,1\}^*$ of length $m \le n$. We want to find a location in the text that best matches the pattern. More precisely, we want to compute

$$\Delta = \min_{i \in \{0,\ldots,n-m\}} d(A[i+1]A[i+2]\cdots A[i+m],\ B[1]B[2]\cdots B[m]),$$

where for two strings s and t of the same length, their "distance" $d(s,t)$ is defined as the number of positions at which the two strings differ. In other words, $d(A[i+1]A[i+2]\cdots A[i+m],\ B[1]B[2]\cdots B[m])$ is equal to $|\{j : A[i+j] \neq B[j]\}|$.

There are known algorithms for computing Δ in $O(n \log n)$ time. Here, we will investigate algorithms for *approximating* Δ that only need to a read a *sublinear* number of characters of the input strings. (Here, we care about the number of characters read, rather than actual run time.) Sublinear algorithms are especially attractive when the input size n is very large.

   (a) [*14 pts*] First consider the simpler problem of approximating the distance between two strings: given two strings s and t of length n, compute a value $\widetilde{d}$ such that $|\widetilde{d} - d(s,t)| \le \varepsilon n$. Describe a simple Monte Carlo algorithm that solves this problem correctly w.h.p. Bound the number of characters read (which should be small, and is a function of n and ε).

   (b) [*14 pts*] Next, consider the original problem. Pick random indices $r_1, \ldots, r_\ell, s_1, \ldots, s_\ell \in \{0, \ldots, \sqrt{n}-1\}$ (you may assume that $\sqrt{n}$ is an integer). Read the following characters:
   - $A[u\sqrt{n} + r_k]$ for all $u \in \{0, \ldots, \sqrt{n}-1\}$ and $k \in \{1, \ldots, \ell\}$; and
   - $B[s_k\sqrt{n} + v]$ for all $v \in \{0, \ldots, \sqrt{n}-1\}$ and $k \in \{1, \ldots, \ell\}$.

   For an appropriate (small) value of ℓ, show that after reading the above $O(\ell\sqrt{n})$ characters, we have enough information to compute a value $\widetilde{\Delta}$ such that $|\widetilde{\Delta} - \Delta| \le \varepsilon n$ w.h.p.

2. [*30 pts*] Recall the problem of approximating the volume of the union of n boxes $B_1, \ldots, B_n$ in $\mathbb{R}^d$. Let V^* denote the volume of the union. Let r be a parameter to be determined. Consider the following variant of the algorithm from class (based on Karp and Luby's technique):

```
0. count = 0, Λ = Σ_{i=1}^{n} vol(B_i)
1. repeat r times {
2.     pick a random i ∈ {1, ..., n} with probability vol(B_i)/Λ
3.     pick a uniformly random point z ∈ B_i
4.     repeat {
5.         increment count
6.         pick a uniformly random j ∈ {1, ..., n}
7.     } until z ∈ B_j
   }
```

(a) [*2 pts*] Let $H_i^{(k)}$ be the set of all points in B_i that are in exactly k of the input boxes. What is $\sum_{i=1}^{n}\sum_{k=1}^{n} \mathrm{vol}(H_i^{(k)})/k$?

(b) [*8 pts*] For each $\ell = 1, \ldots, r$, let X_ℓ be the number of times *count* is incremented during iteration ℓ of the outer loop in lines 1–7. Prove that $\mathbb{E}[X_\ell] = nV^*/\Lambda$ for every ℓ.

(c) [*4 pts*] Prove that $X_\ell \le O(n \log n)$ for all ℓ w.h.p.

(d) [*12 pts*] Assume that a factor-2 approximation V_0 of V^* is given (with $V^* \le V_0 \le 2V^*$). Describe how to set the parameter r so that a factor-$(1 \pm \delta)$ approximation $\widetilde{V}$ of V^* (with $(1-\delta)V^* \le \widetilde{V} \le (1+\delta)V^*$) for any given $\delta > 0$ can be computed w.h.p. from the value *count* found by the above algorithm. Analyze the total expected running time, which should be better than the quadratic bound from class.

(e) [*4 pts*] Describe how to remove the assumption that a factor-2 approximation is given.

3. [*26 pts*] Consider n independent random variables $X_1, \ldots, X_n$, where each X_i is geometrically distributed with probability 1/2, i.e., $\Pr[X_i = \ell] = 1/2^\ell$ for each $\ell = 1, 2, \ldots$

(a) [*13 pts*] Prove that $\Pr[\sum_{i=1}^{n} X_i > 2.01n] \le e^{-\Theta(n)}$, by modifying the proof of Chernoff's bound.

(b) [*13 pts*] Prove that $\Pr[\sum_{i=1}^{n} X_i^2 > 6.01n] \le e^{-\Theta(n^\alpha)}$ for some constant $0 < \alpha < 1$. Aim for the best α.
Note: $\sum_{\ell=1}^{\infty} \ell^2/2^\ell = 6$.
Hint: This part could be trickier (adapting Chernoff's proof doesn't immediately seem to work). One (not necessarily best) approach is to apply Hoeffding's bound to $\sum_{i=1}^{n} \min\{X_i^2, b\}$ for some fixed value b...

4. [*16 pts*] Consider a random binary string $s = a_1 \cdots a_n \in \{0,1\}^*$, where each a_i is chosen from $\{0,1\}$ uniformly and independently.

Given an even number k, let P be the number of length-k palindromic substrings of s, i.e., $P = |\{i : a_{i+1} \cdots a_{i+k} \text{ is a palindrome}\}|$.

(a) [*4 pts*] What is $\mu = \mathbb{E}[P]$ (as a function of n and k)?

(b) [*12 pts*] Bound $\Pr[|P - \mu| > t]$ (in terms of n, k, and t) by using Azuma's inequality and the method of bounded differences.