Problem Set 2 CS265, Winter 2021-2022

Due: January 19 (Wednesday) at 11am (Pacific Time)

Please follow the homework policies on the course website.

---

1. **(7 pt.) [Gotta catch 'em all?]**

   Let M be an unknown set of molecules of size $|M| = n$ that are all present in a liquid solution. You want to identify the set M using an experiment. One run of your experiment on the solution can identify and output a uniformly random molecule from the set M. You can conduct multiple experiments on this solution. Assume that the result of each experiment is independent of the others.

   (a) **(1 pt.)** Give the best lower bound you can to the expected number of experiments you must run to identify all the n distinct molecules in M. To identify a molecule, it must appear as the output of at least one experiment. Use big Omega notation to report a simple answer.

   (b) **(4 pt.)** Suppose the set M of molecules is structured enough for the following to be possible. If you know any $0.99n$ of the items in M, you can infer the other $0.01n$. Thus you will stop conducting experiments after identifying $0.99n$ distinct molecules. What is the expected number of experiments? Show your work and use big O notation to report a simple answer.

   [**HINT:** *Linearity of expectation is still your friend.*]

   (c) **(2 pt.)** Solution A contains molecules from a set S of size n. However, S has no helpful structure. To learn S from Solution A, you use Strategy 1.

   Strategy 1: Run experiments on Solution A until each of the n molecules of S has been observed as the output of an experiment at least once.

   On the other hand, Solution B contains molecules from a different and larger set S'. $|S'| = 10n$ and one can infer the set S from S'. Moreover, S' *is* nicely structured. You can infer S' from any of its subsets of size $9.9n$. To learn S from Solution B, you use Strategy 2.

   Strategy 2:

   i. Run experiments on Solution B until at least $9.9n$ distinct molecules have appeared as the output of an experiment at least once each.

   ii. Infer the set S from the subset of S' of size $9.9n$ you now know.

   Your goal is to find the set S and minimize the expected number of experiments you need to run. Do you choose Strategy 1 or 2?[1] Provide a sentence or two of justification for your answer.

   (d) **(0 pt.) [Optional: this won't be graded.]**

   Can you strengthen the argument for your answer to part (c) by coming up with high probability statements for parts (a) and (b) rather than statements in expectation? [**HINT:** *Try to compute an appropriate variance and use Chebyshev's inequality*]

[1] This scenario is less contrived than you might think, and features in systems where information is stored in DNA. In these systems, enlarging the set from S to S' corresponds to using an error-correcting-code to add redundancy.

2. **(12 pt.) [Tightness of Markov's and Chebyshev's Inequalities]**

   (a) **(4 pt.)** Show that Markov's inequality is tight. Specifically, for each value $c > 1$, describe a distribution D_c supported on non-negative real numbers such that if the random variable X is drawn according to D_c then 1) $\mathbb{E}[X] > 0$ and 2) $\Pr[X \geq c\mathbb{E}[X]] = 1/c$.

   (b) **(4 pt.)**Show that Chebyshev's inequality is tight. Specifically, for each value $c > 1$, describe a distribution D_c supported on real numbers such that if the random variable X is drawn according to D_c then 1) $\mathbb{E}[X] = 0$ and $\mathrm{Var}[X] = 1$ and 2) $\Pr[|X - \mathbb{E}[X]| \geq c\sqrt{\mathrm{Var}[X]}] = 1/c^2$.

   (c) **(4 pt.)** [One-sided version of Chebyshev's Inequality] Prove a one-sided bound on the distribution of a random variable X given its variance. That is, if $\mathrm{Var}[X] = 1$, what the best upper bound on $\Pr[X - \mathbb{E}[X] \geq t]$? Give your answer in terms of t. Prove your bound (a) is true and (b) is tight by coming up with a variable X with distribution D_t and variance 1 for which $\Pr[X - \mathbb{E}[X] \geq t]$ equals your answer.

3. **(0 pt.) [This whole problem is optional and will not be graded.]** In this problem, you'll analyze a different primality test than we saw in class. This one is called the *Agrawal-Biswas Primality test.*

   Given a degree d polynomial $p(x)$ with integer coefficients, for any polynomial $q(x)$ with integer coefficients, we say $q(x) \equiv t(x) \mod (p(x), n)$ if there exists some polynomial $s(x)$ such that $q(x) = s(x) \cdot p(x) + t(x) \mod n$. (Here, we say that $\sum_i c_i x^i = \sum_i c'_i x^i \mod n$ if and only if $c_i = c'_i \mod n$ for all i.) For example, $x^5 + 6x^4 + 3x + 1 \equiv 3x + 1 \mod (x^2 + x, 5)$, since $(x^3)(x^2 + x) + (3x + 1) = x^5 + x^4 + 3x + 1 \equiv x^5 + 6x^4 + 3x + 1 \mod 5$.

   Agrawal-Biswas Primality Test.
   Given n:

   - If n is divisible by 2,3,5,7,11, or 13, or is a perfect power (i.e. $n = c^r$ for integers c and r) then output **composite**.
   - Set d to be the smallest integer greater than $\log n$, and choose a random degree d polynomial with leading coefficient 1:

     $$r(x) = x^d + c_{d-1}x^{d-1} + \ldots + c_1 x + c_0,$$

     by choosing each coefficient c_i uniformly at random from $\{0, 1, \ldots, n-1\}$.
   - If $(x+1)^n \equiv x^n + 1 \mod (r(x), n)$ then output **prime**, else output **composite**.

   Consider the following theorem (you can assume this if you like, or for even more optional work, try to prove it!):

   Theorem 1 (Polynomial version of Fermat's little theorem)**.**

   - *If n is prime, then for any integer a, $(x - a)^n = x^n - a \mod n$.*
   - *If n is not prime and is not a power of a prime, then for any a s.t. $gcd(a, n) = 1$ and any prime factor p of n, $(x - a)^n \neq x^n - a \mod p$.*

First, show that if n is prime, then the Agrawal-Biswas primality test will always return **prime**.

Now, we will prove that if n is composite, the probability over random choices of $r(x)$ that the algorithm successfully finds a witness to the compositeness of n (and hence returns **composite**) is at least $\frac{1}{4d}$.

(a) Using the polynomial version of Fermat's Little Theorem, and the fact that, for prime q, every polynomial over $\mathbb{Z}_q$ that has leading coefficient 1 (i.e. that is "monic") has a unique factorization into irreducible monic polynomials, prove that the number of irreducible degree d factors that the polynomial $(x+1)^n - (x^n+1)$ has over $\mathbb{Z}_p$ is at most n/d, where p is any prime factor of n. (A polynomial is irreducible if it cannot be factored, for example $x^2+1 = (x+1)(x+1) \mod 2$ is not irreducible over $\mathbb{Z}_2$, but x^2+1 is irreducible over $\mathbb{Z}_3$.)

[**HINT:** *Even though this question sounds complicated, the proof is just one line...*]

(b) Let $f(d,p)$ denote the number of irreducible monic degree d polynomials over $\mathbb{Z}_p$. Prove that if n is composite, and not a power of a prime, the probability that $r(x)$ is a witness to the compositeness of n is at least $\frac{f(d,p)-n/d}{p^d}$, where p is a prime factor of n.

[**HINT:** p^d *is the total number of monic degree* d *polynomials over* $\mathbb{Z}_p$.]

(c) Now complete the proof, and prove that the algorithm succeeds with probability at least $1/(4d)$, leveraging the fact that the number of irreducible monic polynomials of degree d over $\mathbb{Z}_p$ is at least $p^d/d - p^{d/2}$. (You should be able to prove a much better bound, though $1/4d$ is fine.)

[**HINT:** *You will also need to leverage the fact that we chose* $d > \log n$ *and also explicitly made sure that* n *has no prime factors less than* 17.]