Problem Set 3 CS265, Autumn 2022-23
Due: October 21 (Friday)

Please follow the homework policies on the course website.

---

1. **(11 pt.) Aggregating Guesses**

   In this problem, we'll consider several different settings where we are aggregating a large number of noisy, unbiased estimates. Suppose a class has n students. Each student is asked to estimate the current temperature. Assume that they each provide independent, unbiased estimates, with X_i denoting the ith student's guess. Let v_i denote $\mathrm{Var}[X_i]$.

   (a) **(4 pt.)** Suppose we know each of the v_i's and decide to compute a weighted combination $Z = \sum_i w_i X_i$, where the weights $w_i \geq 0$ are chosen so as to minimize the variance of Z, subject to $\sum_i w_i = 1$. What are those optimal weights as a function of the v_i's, and roughly how accurate will Z be? Please give an answer of the form: "with probability at least 0.9, Z will be within *blah* of the true temperature, where *blah* is a function of the v_i's.

   (b) **(5 pt.)** For this part, assume each X_i is drawn from a normal (Gaussian) distribution, whose mean is the true temperature, and whose variance is 1. Roughly how accurate should we expect the *median* of the n guesses to be? As above, please give an answer of the form: "with probability at least 0.9, the median of the X_i's will be within *blah* of the true temperature," where *blah* is a function of n. Your value of *blah* should be accurate up to a constant factor and use big-Oh notation, for example $O(1/n^{3/4})$ or something like that.

   [**HINT:** *The following basic fact about a Gaussian should be helpful, and is the only property of a Gaussian that you will need: if Y is a Gaussian with mean μ and variance 1, for any $\epsilon \in (0, 1/2)$ $\Pr[Y < \mu - \epsilon] = \Pr[Y > \mu + \epsilon] < 1/2 - 0.3\epsilon$.*]

   (c) **(2 pt.)** Answer the same question as above for the *mean* of the n values X_i. How do your answers compare?

   (d) **(0 pt.) [Optional: This is a research-level problem.]** As above, suppose each X_i is independently drawn from a normal distribution whose mean is the true temperature, and variance v_i. Assume you know the (multi)set of the v_i's, but you don't know which variance corresponds to which guess. How well should you expect to do, and is there an efficient algorithm that achieves this?

   (e) **(0 pt.) [Optional: This is a research-level problem.]** Suppose we are in the setting above, but don't know anything about the variances. What is a near-optimal algorithm, and how well will it do, as a function of the (unknown) list of variances $v_1, \ldots$.?

   [**HINT:** *Note that if two X_i's are identical (or super, super close) then we know that two of the variances are 0 (or really, really small), and hence either of those X_i's would give an extremely accurate guess, no matter what the other $n - 2$ guesses are...*]

2. **(11 pt.) Concentration without Independence**

   A computer system has n different failure modes, each of which happens with a small probability. Fortunately, the system is designed to be sufficiently robust in the following sense: as

long as less than half of the failures occur, things are fine; otherwise, a large-scale crash will happen. We want to make sure that the probability of crashing is small enough.

To model the above scenario, we define n Bernoulli random variables $X_1, \ldots, X_n$. Each X_i is the indicator of the i-th failure mode, i.e., $X_i = 1$ if failure i occurs and $X_i = 0$ otherwise. Our goal is to upper bound the probability $\Pr\left[\sum_{i=1}^n X_i \geq n/2\right]$.

(a) **(2 pt.)** Let's first assume that the n failure events are independent and the probability of each failure is at most $1/3$. Formally, we have:

Assumption 1. $\Pr[X_i = 1] \leq 1/3$ *for every* $i \in [n]$ *and* $X_1, \ldots, X_n$ *are independent.*

Prove that under Assumption 1, for some constant $C > 0$ that does not depend on n,

$$\Pr\left[\sum_{i=1}^n X_i \geq n/2\right] \leq e^{-Cn}. \tag{1}$$

Thus, the probability of a crash is exponentially small in n.

[**HINT:** *Feel free to use (without proof) any of the Chernoff bounds in lecture note #5 (including Theorem 2 and Corollaries 5 and 6) and also the inequality* $\frac{e^\delta}{(1+\delta)^{1+\delta}} \leq e^{-\delta^2/3}$ *for* $\delta \in [0, 1]$.]

(b) **(1 pt.)** Now we decide that Assumption 1 is too unrealistic, since many of the failure modes are known to be strongly correlated. Show that only assuming $\Pr[X_i = 1] \leq 1/3$ (and not the independence), the probability of crashing can be as large as $\Omega(1)$. In particular, prove that for any $n \geq 1$, there exist random variables $X_1, \ldots, X_n$ that satisfy: (1) $\Pr[X_i = 1] \leq 1/3$ for every $i \in [n]$; (2) $\Pr\left[\sum_{i=1}^n X_i \geq n/2\right] \geq 1/3$.

(c) **(2 pt.)** Let's try the following relaxation of Assumption 1, which states that the probability for k different failures to occur simultaneously is exponentially small in k:

Assumption 2. *For any* $S \subseteq [n]$, $\Pr\left[X_i = 1 \text{ for all } i \in S\right] \leq (1/3)^{|S|}$.

Show that Assumption 2 is strictly weaker than Assumption 1 by proving: (1) Assumption 1 implies Assumption 2; (2) the implication on the other direction does not hold, i.e., there exist some n and $X_1, \ldots, X_n$ that satisfy Assumption 2 but not Assumption 1.

[**HINT:** *For (2), there exists a counterexample for* $n = 2$.]

(d) **(6 pt.)** Prove that under Assumption 2, inequality (1) holds for some constant $C > 0$. In your proof, you can appeal to the proof of the Chernoff bounds from lecture videos/notes if you need to write it out verbatim at some point. For example, if you manage to upper bound $\Pr\left[\sum_{i=1}^n X_i \geq n/2\right]$ by an expression involving the moment-generating function of some random variable Y that is the sum of n independent Bernoulli random variables, you can simply say that "the rest of the proof is exactly the proof of Theorem 2 from Lecture #5".

[**HINT:** *Consider independent Bernoulli random variables* $Y_1, \ldots, Y_n$ *with* $Pr[Y_i = 1] = 1/3$ *for each* $i \in [n]$. *For distinct indices* $i, j, \ell \in [n]$, *does* $\mathbb{E}\left[X_i X_j X_\ell\right] \leq \mathbb{E}\left[Y_i Y_j Y_\ell\right]$ *hold? Can you extend your proof of the inequality to the case with repeating indices?*]

[**HINT:** *Let* $X = \sum_{i=1}^n X_i$ *and* $Y = \sum_{i=1}^n Y_i$. *What can we say about* $\mathbb{E}\left[X^k\right]$ *and* $\mathbb{E}\left[Y^k\right]$ *for integer* $k \geq 0$*? Considering the identity* $e^z = \sum_{k=0}^{+\infty} \frac{z^k}{k!}$, *what can we say about* $\mathbb{E}\left[e^{tX}\right]$ *and* $\mathbb{E}\left[e^{tY}\right]$ *for any* $t > 0$*?*]

(e) **(0 pt.) [Optional: this won't be graded.]** Can you construct counterexamples for Part 2b that satisfy *pairwise independence* but have a crashing probability of $\Omega(1/n)$? Formally, prove that there exists $C > 0$ such that for any $n \geq 2$, there exist $X_1, \ldots, X_n$ that satisfy: (1) $\Pr[X_i = 1] \leq 1/3$; (2) X_i and X_j are independent for distinct $i, j \in [n]$; (3) $\Pr\left[\sum_{i=1}^{n} X_i \geq n/2\right] \geq C/n$.

[**NOTE:** *This shows that unlike Chebyshev's inequality, Chernoff bounds do not hold if we only assume pairwise independence.*]

[**HINT:** *Recall pairwise independent hash functions if you have seen them before. You can use the Bertrand-Chebyshev theorem, which states that for any integer $n \geq 1$, there exists a prime number p with $n < p < 2n$.*]

3. **(8 pt.) Processes and CPUs**

Suppose that in a distributed system, we have N CPUs and P processes. Each process is independently and uniformly allocated to a CPU. However, if multiple processes are allocated to the same CPU, the CPU will choose one of them at random to complete; the remaining processes allocated to that CPU will not be completed.

(a) **(2 pt.)** What is expected number of processes that will be completed?

(b) **(6 pt.)** Suppose that $P \geq N$, and denote the total number of completed processes by C. **Use Poissonization** to prove that $\Pr\left[C \leq \frac{N}{2}\right] \leq e^{-\Omega(N)}$.

[**Note:** There may be ways to do this problem that don't involve Poissonization, but we want you to use it to get practice with it. That is, you should prove the statement by analyzing the case when the number of processes is an appropriate Poisson random variable. Don't forget the de-Poissonization step!]

(c) **(0 pt.) [Optional: this won't be graded.]** Let μ be your answer from part (a). Under what conditions on P and N can you use Poissonization to show that $\Pr[C \leq (1-\delta)\mu] \leq e^{-\Omega_\delta(N)}$ for any $\delta > 0$, where the Ω_δ notation means that you are allowed to have constants that depend on δ hidden inside the big-Omega.