(CS1.406) Advanced Algorithms (Spring 2024) Due: 07.05.2024

Problem Set 3

Instructions:

- Discussions amongst the students are not discouraged, but all writeups must be done individually and must include names of all collaborators.
- Referring sources other than the lecture notes is discouraged as solutions to some of the problems can be found easily via a web search. But if you do use an outside source (eg., text books, other lecture notes, any material available online), do mention the same in your writeup. This will not affect your grades. However dishonesty of any sort when caught shall be heavily penalized.
- Be clear in your arguments. Vague arguments shall not be given full credit.

---

1. We have a function $F : \{0, \ldots, n-1\} \mapsto \{0, \ldots, m-1\}$. We know that, for $0 \leqslant x, y \leqslant n-1$, $F((x+y) \mod n) = (F(x) + F(y)) \mod m$. The only way we have for evaluating F is to use a lookup table that stores the values of F . Unfortunately, an Evil Adversary has changed the value of 1/5th of the table entries when we were not looking.

   Describe a simple randomized algorithm that, given an input z, outputs a value that equals $F(z)$ with probability at least $1/2$. Your algorithm should work for every value of z, regardless of what values the Adversary changed. Your algorithm should use as few lookups and as little computation as possible. Suppose we allow you to repeat your initial algorithm three times. What should you do in this case, and what is the probability that your enhanced algorithm returns the correct answer?

2. Let X be a random variable with expectation 0 such that the moment generating function $\mathbb{E}[\exp(t|X|)]$ is finite for some $t \geqslant 0$. We can use the following two kinds of tail inequalities on X.

   - Chernoff bound:

   $$\mathbb{P}[|X| \geqslant \delta] \leqslant \min_{t \geqslant 0} \frac{\mathbb{E}\left[e^{t|X|}\right]}{e^{t\delta}}.$$

   - kth-moment bound:

   $$\mathbb{P}[|X| \geqslant \delta] \leqslant \frac{\mathbb{E}\left[|X^k|\right]}{\delta^k}.$$

   Show that for each δ, there exists a k such that the kth-moment bound is stronger than Chernoff bound. (**Hint:** Use taylor expansion on the moment generating function and use probabilistic method.)

3. We plan to conduct an opinion poll to find out the percentage of people in a community who want its president impeached. Assume that every person answers either yes or no. If the actual fraction of people who want the president impeached is p, we want to find an estimate X of p such that

$$\mathbb{P}\left[|X - p| \leqslant \varepsilon \cdot p\right] > 1 - \delta$$

for a given ε and δ, with $0 < \varepsilon, \delta < 1$. We query N people chosen independently and uniformly at random from the community and output the fraction of them who want the president impeached. How large should N be for our result to be a suitable estimator of p? Use Chernoff bounds, and express N in terms of p, ε, and δ.

4. To improve the probability of success of the randomized min-cut algorithm, it can be run multiple times.

   (a) Consider running the algorithm twice. Determine the number of edge contractions and bound the probability of finding a min-cut.

   (b) Consider the following variation. Starting with a graph with n vertices, first contract the graph down to k vertices using the randomized min-cut algorithm. Make copies of the graph with k vertices, and now run the randomized algorithm on this reduced graph ℓ times, independently. Determine the number of edge contractions and bound the probability of finding a minimum cut.

   (c) Find optimal (or at least near-optimal) values of k and l for the variation in (b) that maximize the probability of finding a minimum cut while using the same number of edge contractions as running the original algorithm twice.

5. Given an n-vertex undirected graph $G = (V, E)$,consider the following method of generating an independent set. Given a permutation σ of the vertices, define a subset $S(\sigma)$ of the vertices as follows: for each vertex i, $i \in S(\sigma)$ if and only if no neighbor j of i precedes i in the permutation σ.

   (a) Show that each $S(\sigma)$ is an independent set in G.

   (b) Suggest a natural randomized algorithm to produce σ for which you can show that the expected cardinality of $S(\sigma)$ is

$$\sum_{i=1}^{n} \frac{1}{d_i + 1}$$

   where d_i denotes the degree of vertex i.

   (c) Prove that G has an independent set of size at least $\sum_{i=1}^{n} \frac{1}{d_i+1}$.

6. We are given a $n \times n$ matrix A all of whose entries are 0 or 1. In addition, we are given a column vector p with n entries, all of which are in the interval $[0, 1]$. We wish to find a column vector q with n entries, all of which are entries in the set $\{0, 1\}$, so as to minimize $\|A(p - q)\|_\infty$. That is, minimize $\max_{i \in [n]}\{(A(p - q))_i\}$. In other words, the column vector q is an integer approximation to column vector p. Derive a bound on $\|A(p - q)\|_\infty$ assuming q were derived from p using randomized rounding.

7. One could consider the following approach for estimating the value of the constant π. Let (X, Y) be a point chosen uniformly at random in a 2×2 square centered at origin. That is, X and Y are chosen independently from a uniform distribution on $[-1, 1]$ (continuous space). A circle of radius 1 centered at $(0, 0)$ lies inside the square, and has area π. Define the random variable Z dependent on X and Y is defined as follows.

$$Z = \begin{cases} 1 & \text{if } \sqrt{X^2 + Y^2} \leqslant 1 \\ 0 & \text{otherwise.} \end{cases}$$

   (a) What is the expected value of Z.

   (b) Let this experiment be run m times by sampling X and Y independently among the runs. Let Z_i be the value of Z in the ith run, and $W = \sum_{i=1}^{m} Z_i$. Using this set-up and the information provided, estimate the value of π as closely[1] as possible.

8. The problem of counting the number of solutions to a knapsack instance can be defined as follows. Given items with integral weights $a_1, a_2, \ldots, a_n > 0$ and integer $b > 0$, find the number of vectors $(x_1, x_2, \ldots, x_n) \in \{0, 1\}^n$ such that $\sum_{i=1}^{n} a_i x_i \leqslant b$. Number b can be thought of as the capacity of the knapsack and each x_i corresponds to whether this item i is picked or not. Counting solutions corresponds to number of different sets of items that can be placed in the bag without exceeding the capacity.

   Consider a Markov chain $X_0, X_1, \ldots$ on vectors $(x_1, x_2, \ldots, x_n)$. Suppose X_j is $(x_1, x_2, \ldots, x_n)$. At each step, Markov chain chooses $i \in [n]$ uniformly at random. If $x_i = 1$, then X_{j+1} is obtained by setting $x_i = 0$. If $x_i = 0$, then X_{j+1} is obtained by setting $x_i = 1$ provided the constraint $\sum_{i=1}^{n} a_i x_i \leqslant b$ is still satisfied after setting x_i to 1. Otherwise, $X_{j+1} = X_j$. Argue that this Markov chain has a uniform stationary distribution whenever $\sum_{i=1}^{n} a_i > b$.

9. Consider a sequence of independent, fair gambling games between two players. In each round a player wins a dollar with probability $\frac{1}{2}$ or loses a dollar with probability $\frac{1}{2}$. We can construct a markov chain such that the state of the system at the time t is the number of dollars won by player 1. If player 1 lost money, this number is negative. Initial state [2] is 0. It is reasonable to assume that there are numbers ℓ_1 and ℓ_2 such that player i (for $i \in \{1, 2\}$) cannot lose more that ℓ_i dollars and thus game ends when the markov chain reaches one of the two states $-\ell_1$ or ℓ_2. At this point, one of the gamblers is ruined. That is, they lost all their money. What is the probability that player 1 wins ℓ_2 dollars before losing ℓ_1 dollars.

10. Consider a random walk on an infinite line. At each step, the position of the particle is one of the integer points. At the next step, it moves to one of the neighbouring integral points with equal probability. Show that the expected distance of the particle from origin after n steps is $\Theta(\sqrt{n})$. (**Hint:** If $\mathbb{E}[|X|]$ is hard to analyze, one could possibly compute $\mathbb{E}\left[X^2\right]$ and then obtain $\mathbb{E}[|X|]$ using Jensen's inequality).

11. Given a set S of n elements drawn from a totally ordered universe, the median of S is an element m of S such that at least $\lfloor n/2 \rfloor$ elements in S are less than or equal to m and at least $\lfloor n/2 \rfloor + 1$ elements in S are greater than or equal to m. If the elements in S are distinct, then m is the $(\lceil n/2 \rceil)$th element

[1] closeness in terms of a parameter ε and m

[2] That is, both the players start with 0 dollars.

in the sorted order of S. The median can be easily found deterministically in $O(n \log n)$ steps by sorting, and there is a relatively complex deterministic algorithm that computes the median in $O(n)$ time.

The main idea of the algorithm involves sampling. The goal is to find two elements that are close together in the sorted order of S and that have the median lie between them. Specifically, we seek two elements $d, u \in S$ such that:

- $d \leqslant m \leqslant u$ (the median is between d and u); and
- for $C = s \in S \mid d \leqslant s \leqslant u$, $|C| = o(n/\log n)$ (the total number of elements between d and u is small).

Algorithm 1: Randomized Median Algorithm

Data: A set S of n elements over a totally ordered universe
Result: The median element of S, denoted by m

1 Pick a (multi-)set R of $\lceil n^{3/4} \rceil$ elements in S, chosen independently and uniformly at random with replacement; Sort the set R;
2 Let d be the $\left(\lfloor \frac{n^{3/4}}{2} - \sqrt{n} \rfloor\right)$th smallest element in the sorted set R;
3 Let u be the $\left(\lfloor \frac{n^{3/4}}{2} + \sqrt{n} \rfloor\right)$th smallest element in the sorted set R;
4 By comparing every element in S to d and u, compute the set $C = s \in S \mid d \leqslant s \leqslant u$ and the numbers $\ell_d = |x \in S : x < d|$ and $\ell_u = |x \in S : x > u|$;
5 If $\ell_d > n/2$ or $\ell_u > n/2$ then FAIL;
6 If $|C| \leqslant 4n^{3/4}$ then sort the set C, otherwise FAIL;
7 Output the $(\lfloor n/2 \rfloor - \ell_d + 1)$th element in the sorted order of C;

Based on this discussion, answer the following questions.

(a) Argue that the afore mentioned algorithm terminates in linear time, and if it does not output FAIL it outputs the correct median element of the input set S.

(b) Let $\mathcal{E}_1, \mathcal{E}_2$ and $\mathcal{E}_3$ be defined as follows.

$$\mathcal{E}_1 : Y_1 = |r \in R \mid r \leqslant m| < \frac{n^{3/4}}{2} - \sqrt{n}$$
$$\mathcal{E}_2 : Y_2 = |r \in R \mid r \geqslant m| < \frac{n^{3/4}}{2} - \sqrt{n}$$
$$\mathcal{E}_3 : |C| > 4n^{3/4}$$

Show that the randomized median algorithm fails if at least one of $\mathcal{E}_1, \mathcal{E}_2$, or $\mathcal{E}_3$ occurs.

(c) Using the definitions of events as above show the following.

i. $\mathbb{P}[\mathcal{E}_1], \mathbb{P}[\mathcal{E}_2] \leqslant \frac{n^{-1/4}}{4}$ and

ii. $\mathbb{P}[\mathcal{E}_3] \leqslant \frac{n^{-1/4}}{2}$.

(d) Put all the parts above together and show that the randomized median algorithm fails with a probability of at most $n^{-1/4}$.

1. We have a function $F : \{0, \ldots, n-1\} \mapsto \{0, \ldots, m-1\}$. We know that, for $0 \leqslant x, y \leqslant n-1$, $F((x+y) \bmod n) = (F(x) + F(y)) \bmod m$. The only way we have for evaluating F is to use a lookup table that stores the values of F. Unfortunately, an Evil Adversary has changed the value of 1/5th of the table entries when we were not looking.

   Describe a simple randomized algorithm that, given an input z, outputs a value that equals $F(z)$ with probability at least 1/2. Your algorithm should work for every value of z, regardless of what values the Adversary changed. Your algorithm should use as few lookups and as little computation as possible. Suppose we allow you to repeat your initial algorithm three times. What should you do in this case, and what is the probability that your enhanced algorithm returns the correct answer?

Choose x uniformly at random from the set $[n-1]$.

Let $y = z - x$.

We output $[F(x) + F(y)] \bmod m$ as our answer.

This answer would be wrong when either $F(x)$ is wrong or when $F(y)$ is wrong (or both).

We do not consider the case when the answer turns out to be right when both $F(x)$ and $F(y)$ are wrong.

$\Pr(F(x) \text{ is wrong}) = \frac{1}{5}$; $\Pr(F(y) \text{ is wrong}) = \frac{1}{5}$

$\therefore \Pr(F(x) \text{ or } F(y) \text{ is wrong}) \leq \frac{1}{5} + \frac{1}{5} = \frac{2}{5}$ (By union bound)

$\therefore$ Error probability $\leq 2/5$

This algorithm uses only 2 lookups and has success probability of at least $\frac{3}{5}$ which is greater than $\frac{1}{2}$.

If we repeat this algorithm 3 times, and take the majority vote if it exist or select any one of the three test's result as an output to the algorithm.

This algorithm may be wrong only when

1) Majority vote doesn't exist

2) Majority vote is incorrect

Let α be the probability that a single test errs.

The probability of the algorithm erring will then be:

Case 1: Majority vote doesn't exist

Pr(2 tests being incorrect) × Pr(correct test not being picked)

+ Pr(all three test being incorrect)

$$= \binom{3}{2} \alpha^2(1-\alpha) \times \frac{2}{3} + \alpha^3$$

$2+8 = 32$

$$2\alpha^2(1-\alpha) + \alpha^3$$

since $\alpha \leq \frac{2}{5} \Rightarrow 2\alpha^2(1-\alpha) + \alpha^3 \leq 2 \cdot \frac{4}{25} \times \frac{3}{5} + \frac{8}{125} = \frac{32}{125}$

$$\frac{32}{125} = 0.256$$

Case 2: Majority vote is incorrect:

Pr(2 test being incorrect) + Pr(3 test being incorrect)

$$= \binom{3}{2} \times \alpha^2(1-\alpha) + \alpha^3 \leq 3 \times \frac{4}{25} \times \frac{3}{5} + \frac{8}{125} = \frac{44}{125} = 0.352$$

2. Let X be a random variable with expectation 0 such that the moment generating function $\mathbb{E}[\exp(t|X|)]$ is finite for some $t \geqslant 0$. We can use the following two kinds of tail inequalities on X.

   - Chernoff bound:

$$\mathbb{P}[|X| \geqslant \delta] \leqslant \min_{t \geqslant 0} \frac{\mathbb{E}\left[e^{t|X|}\right]}{e^{t\delta}}.$$

   - kth-moment bound:

$$\mathbb{P}[|X| \geqslant \delta] \leqslant \frac{\mathbb{E}\left[|X^k|\right]}{\delta^k}.$$

   Show that for each δ, there exists a k such that the kth-moment bound is stronger than Chernoff bound. (**Hint:** Use taylor expansion on the moment generating function and use probabilistic method.)

We have

$$E(e^{tX}) = E\left[\sum_{n\geq 0} \frac{t^n x^n}{n!}\right] = \sum_{n\geq 0} \frac{t^n E(x^n)}{n!}$$

$$\therefore \frac{E(e^{tX})}{e^{t\delta}} = \frac{\sum_{n\geq 0} \frac{t^n E(x^n)}{n!}}{\sum_{n\geq 0} \frac{s^n \delta^n}{n!}} = \lim_{n\to\infty} \frac{1 + tE(x) + \cdots + \frac{t^n E(x^n)}{n!}}{1 + t\delta + \cdots + \frac{t^n \delta^n}{n!}}$$

We have cauchy's third inequality as follows:

$$\frac{a_1 + \cdots + a_n}{b_1 + \cdots + b_n} \geq \min_{k\leq n} \frac{a_k}{b_k}$$

Using cauchy's third inequality, we get:

$$\frac{1 + tE(x) + \cdots + \frac{t^n E(x^n)}{n!}}{1 + t\delta + \cdots + \frac{t^n \delta^n}{n!}} \geq \min_{k\leq n} \frac{t^k E(x^k) k!}{k!\, t^k \delta^k} = \min_{k\leq n} \frac{E(x^k)}{\delta^k}$$

∴ By taking limit, we get:

$$\frac{E(e^{tx})}{e^{t\delta}} \geq \inf_{K \in \mathbb{N}} \frac{E(x^K)}{\delta^K}$$

3. We plan to conduct an opinion poll to find out the percentage of people in a community who want its president impeached. Assume that every person answers either yes or no. If the actual fraction of people who want the president impeached is p, we want to find an estimate X of p such that

$$\mathbb{P}\left[|X - p| \leqslant \varepsilon \cdot p\right] > 1 - \delta$$

for a given ε and δ, with $0 < \varepsilon, \delta < 1$. We query N people chosen independently and uniformly at random from the community and output the fraction of them who want the president impeached. How large should N be for our result to be a suitable estimator of p? Use Chernoff bounds, and express N in terms of p, ε, and δ.

Let $X_i = \begin{cases} 1 & \text{if the } i^{th} \text{ person says yes} \\ 0 & \text{o/w} \end{cases}$

$Pr(X_i = 1)$ = Probability of a randomly selected person says yes $= p$.

$\therefore Pr(X_i = 1) = p$.

$\therefore E(X_i) = 1 \cdot p + 0 \cdot p = p$

Let $X = \frac{\sum_{i=1}^{N} X_i}{N}$

$$E(X) = E\left[\frac{1}{N}\sum_{i=1}^{N} X_i\right] = \frac{1}{N}\sum_{i=1}^{N} E(X_i) = \frac{1}{N}\sum_{i=1}^{N} p = p$$

We have chernoff bound as follows:

$$Pr[\,|X - \mu| \geq \delta\mu\,] \leq 2e^{-\frac{\mu\delta^2}{3}}$$

$$\therefore \quad Pr[\,|x-p| > \varepsilon p\,]$$

$$\Rightarrow Pr\left[\,\left|\sum_{i=1}^{N} x_i - Np\right| > \varepsilon Np\,\right] < 2e^{\frac{-(Np)\varepsilon^2}{3}}$$

$$\therefore \quad 2e^{\frac{-(Np)\varepsilon^2}{3}} < \delta$$

$$\Rightarrow \quad -\frac{(Np)\varepsilon^2}{3} < \log\left(\frac{\delta}{2}\right)$$

$$\Rightarrow \quad (Np)\varepsilon^2 > 3\log\left(\frac{2}{\delta}\right)$$

$$\Rightarrow \quad N > \frac{3\log\left(\frac{2}{\delta}\right)}{p\,\varepsilon^2}$$

4. To improve the probability of success of the randomized min-cut algorithm, it can be run multiple times.

   (a) Consider running the algorithm twice. Determine the number of edge contractions and bound the probability of finding a min-cut.

   (b) Consider the following variation. Starting with a graph with n vertices, first contract the graph down to k vertices using the randomized min-cut algorithm. Make copies of the graph with k vertices, and now run the randomized algorithm on this reduced graph ℓ times, independently. Determine the number of edge contractions and bound the probability of finding a minimum cut.

   (c) Find optimal (or at least near-optimal) values of k and l for the variation in (b) that maximize the probability of finding a minimum cut while using the same number of edge contractions as running the original algorithm twice.

(a) The no. of edge contraction should be $2(n-2)$.

The prob. that it finds a min-cut is $(1-p)^2$

where $p = \frac{2}{n(n-1)}$.

(b) The no. of edge contractions should be

$$(n-k) + \ell(k-2)$$

$$p = \frac{k(k-1)}{n(n-1)} \cdot \left(1 - \left(1 - \frac{2}{k(k-1)}\right)^{\ell}\right) \approx \frac{k^2}{n^2}\left(1 - e^{-2\ell/k^2}\right)$$

(c) We know the condition is that the no. of edge contractions is $(n-k) + \ell(k-2) = 2(n-2)$.

$\therefore \ell = \frac{n+k-4}{k-2}$ which is roughly equal to

$$\frac{k^2}{n^2}\left(1 - e^{-\frac{(n+k)}{3}}\right).$$

This maximized when $K = n^{1/3}$ as

for $K \geq n^{1/3}$, since $1 - e^{-x} \leq x \quad \forall x \geq 0$

this is at most $\frac{K^2}{n^2}\left(\frac{n+K}{K^3}\right) = \frac{n+K}{Kn^2}$. Similarly,

if $K \leq n^{1/3}$, then

$\left(1 - e^{-\frac{(n+K)}{K^3}}\right)$ is at least a constant bounded away from zero, and hence upto a constant factor, the prob. of success is $O(n^{2/3}/n^2) = \Omega(n^{-4/3})$

5. Given an n-vertex undirected graph $G = (V, E)$,consider the following method of generating an independent set. Given a permutation σ of the vertices, define a subset $S(\sigma)$ of the vertices as follows: for each vertex i, $i \in S(\sigma)$ if and only if no neighbor j of i precedes i in the permutation σ.

   (a) Show that each $S(\sigma)$ is an independent set in G.

   (b) Suggest a natural randomized algorithm to produce σ for which you can show that the expected cardinality of $S(\sigma)$ is

$$\sum_{i=1}^{n} \frac{1}{d_i + 1}$$

   where d_i denotes the degree of vertex i.

   (c) Prove that G has an independent set of size at least $\sum_{i=1}^{n} \frac{1}{d_i+1}$.

(a) Denote the permutation σ as

$$\sigma = v_1, v_2, \cdots, v_i, \cdots, v_n.$$

where n is the # of vertex.

And $S(\sigma) = d_1, d_2, \cdots, d_K$, where K is the size of $S(\sigma)$.

Clearly, $d_1 = v_1 \in S(\sigma)$ as the first element.

Let $d_i \in S(\sigma)$. There's no connection with the d_j where $j < i$, because none of its neighbours precede before it.

And any d_j where $j > i$ has no connection with d_i for the same reason. Therefore, the vertices in $S(\sigma)$ form an independence set in G.

(b) The expectation of v_i in the subset $S(\sigma)$ is $E(v_i) = \frac{1}{d_i+1}$

for the reason of that v_i need to occur first within its d_i neighbours and itself, so the probability of that is $\frac{1}{d_i+1}$.

∴ From linearity of Expectation, by considering all vertices in $S(\sigma)$.

$$E(S(\sigma)) = \sum_{i=1}^{n} \frac{1}{d_i+1}$$

(c) From lemma 6.2 in Mitzenmacher:

Lemma 6.2: *Suppose we have a probability space $\mathcal{S}$ and a random variable X defined on $\mathcal{S}$ such that $\mathbf{E}[X] = \mu$. Then $\Pr(X \geq \mu) > 0$ and $\Pr(X \leq \mu) > 0$.*

Proof: We have

$$\mu = \mathbf{E}[X] = \sum_x x \Pr(X = x),$$

where the summation ranges over all values in the range of X. If $\Pr(X \geq \mu) = 0$, then

$$\mu = \sum_x x \Pr(X = x) = \sum_{x<\mu} x \Pr(X = x) < \sum_{x<\mu} \mu \Pr(X = x) = \mu,$$

giving a contradiction. Similarly, if $\Pr(X \leq \mu) = 0$ then

$$\mu = \sum_x x \Pr(X = x) = \sum_{x>\mu} x \Pr(X = x) > \sum_{x>\mu} \mu \Pr(X = x) = \mu,$$

again yielding a contradiction. ■

since expectation is $\sum_{i=1}^{n} \frac{1}{d_i+1}$,

$$\Pr\left(\text{Independent set of size} \geq \sum_{i=1}^{n} \frac{1}{d_i+1}\right) > 0.$$

6. We are given a $n \times n$ matrix A all of whose entries are 0 or 1. In addition, we are given a column vector p with n entries, all of which are in the interval $[0, 1]$. We wish to find a column vector q with n entries, all of which are entries in the set $\{0, 1\}$, so as to minimize $\|A(p - q)\|_\infty$. That is, minimize $\max_{i \in [n]}\{(A(p - q))_i\}$. In other words, the column vector q is an integer approximation to column vector p. Derive a bound on $\|A(p - q)\|_\infty$ assuming q were derived from p using randomized rounding.

since q_j is 1 w.p. p_j and 0 w.p. $1-p_j$,

the expected value of q_j is p_j and therefore

$$E(p_j - q_j) = 0.$$

Applying Hoeffding's Inequality to each entry of $A(p-q)$, we get that for each i, the probability that $(A(p-q))_i$ deviates from its expected value more than t is bounded by:

$$Pr(\,|(A(p-q))_i - E[(A(p-q))_i]| \geq t) \leq 2\exp\left(\frac{-2t^2}{\sum_{j=1}^{n}(b_j - a_j)^2}\right)$$

where a_j and b_j are the minimum and maximum possible values of $p_j - q_j$, which are -1 and 1 respectively

since $p_j \in [0,1]$ and $q_j \in \{0,1\}$.

$$\therefore \sum_{j=1}^{n}(b_j - a_j)^2 = \sum_{j=1}^{n} 4 = 4n.$$

$$\therefore Pr(\,|(A(p-q))_i - E[(A(p-q))_i]| \geq t) \leq 2\exp\left(\frac{-t^2}{2n}\right)$$

By taking a union bound over all i's, we get

$$\Pr\left(\|A(p-q)\|_\infty > t \right) \leq n \cdot 2\exp\left(-\frac{t^2}{2n}\right)$$

setting δ as the RHS, we get:

$$\delta = n \cdot 2\exp\left(-\frac{t^2}{2n}\right)$$

$$\log\left(\frac{2n}{\delta}\right) = \frac{t^2}{2n}$$

$$\therefore \quad t = \sqrt{2n \log\left(\frac{2n}{\delta}\right)}$$

7. One could consider the following approach for estimating the value of the constant π. Let (X, Y) be a point chosen uniformly at random in a 2×2 square centered at origin. That is, X and Y are chosen independently from a uniform distribution on $[-1, 1]$ (continuous space). A circle of radius 1 centered at $(0, 0)$ lies inside the square, and has area π. Define the random variable Z dependent on X and Y is defined as follows.

$$Z = \begin{cases} 1 & \text{if } \sqrt{X^2 + Y^2} \leqslant 1 \\ 0 & \text{otherwise.} \end{cases}$$

(a) What is the expected value of Z.

(b) Let this experiment be run m times by sampling X and Y independently among the runs. Let Z_i be the value of Z in the ith run, and $W = \sum_{i=1}^{m} Z_i$. Using this set-up and the information provided, estimate the value of π as closely[1] as possible.

(a) $\Pr(Z=1) = \Pr(\text{the point within the circle of radius } \frac{1}{2})$

$= \dfrac{\text{Area of circle with radius } 1/2}{\text{Area of unit square}} = \dfrac{\pi(\frac{1}{2})^2}{1} = \dfrac{\pi}{4}$.

$\therefore E(Z) = 1 \cdot \Pr(Z=1) + 0 \cdot \Pr(Z=0)$

$= 1 \cdot \frac{\pi}{4} = \frac{\pi}{4}$.

(b) We have,

$$W = \sum_{i=1}^{m} Z_i$$

$$E(W) = E\left(\sum_{i=1}^{m} Z_i\right) = \sum_{i=1}^{m} E(Z_i) = \sum_{i=1}^{m} \frac{\pi}{4} = \frac{m\pi}{4}$$

To give a (ε, δ) estimate of π, we need the value of $\frac{4W}{m}$.

$$\therefore \Pr\left(\left|W - \frac{m\pi}{4}\right| > \varepsilon \cdot \left(\frac{m\pi}{4}\right)\right) < 2e^{-\left(\frac{m\pi}{4}\right)\varepsilon^2/3}$$

$$\therefore \quad \delta > 2e^{-\frac{(m\pi)}{4}\varepsilon^2/3}$$

$$\Rightarrow \log\left(\frac{\delta}{2}\right) > -\frac{(m\pi)}{4}\varepsilon^2/3$$

$$\Rightarrow \quad m > \frac{12}{\pi\varepsilon^2} \cdot \log\left(\frac{2}{\delta}\right)$$

8. The problem of counting the number of solutions to a knapsack instance can be defined as follows. Given items with integral weights $a_1, a_2, \ldots, a_n > 0$ and integer $b > 0$, find the number of vectors $(x_1, x_2, \ldots, x_n) \in \{0,1\}^n$ such that $\sum_{i=1}^{n} a_i x_i \leqslant b$. Number b can be thought of as the capacity of the knapsack and each x_i corresponds to whether this item i is picked or not. Counting solutions corresponds to number of different sets of items that can be placed in the bag without exceeding the capacity.

   Consider a Markov chain $X_0, X_1, \ldots$ on vectors $(x_1, x_2, \ldots, x_n)$. Suppose X_j is $(x_1, x_2, \ldots, x_n)$. At each step, Markov chain chooses $i \in [n]$ uniformly at random. If $x_i = 1$, then X_{j+1} is obtained by setting $x_i = 0$. If $x_i = 0$, then X_{j+1} is obtained by setting $x_i = 1$ provided the constraint $\sum_{i=1}^{n} a_i x_i \leqslant b$ is still satisfied after setting x_i to 1. Otherwise, $X_{j+1} = X_j$. Argue that this Markov chain has a uniform stationary distribution whenever $\sum_{i=1}^{n} a_i > b$.

we first need to show the Markov chain is irreducible and aperiodic over all the states of all valid solutions of the Knapsack problem. Once we show this, it would mean the chain is ergodic and therefore has a unique stationary distribution according to the fundamental theorem of Markov chain taught in class.

→ For any solution x, there is a positive probability of going back down to the all zero vector by zeroing out the non-zero x_i's one-by-one. On the other hand, if x is a solution, then it is possible to reach x starting from the all zero vector. Therefore, the chain is irreducible.

→ Given $\sum_{i=1}^{n} a_i > b$, then there must exist $j \in [n]$ and a vector $x = (x_1, x_2, \ldots, x_j = 0, \ldots, x_n) \in \{0,1\}^n$ s.t. $\sum_{i=1}^{n} a_i x_i \leq b$ but $\sum_{i=1}^{n} a_i x_i + a_j > b$.

This means the self-loop probability in the Markov chain is $P_{x,x} > 0$. Therefore, the chain is aperiodic.

Now, let M be the no. of solutions and let x and y be two solutions that differ by one bit.

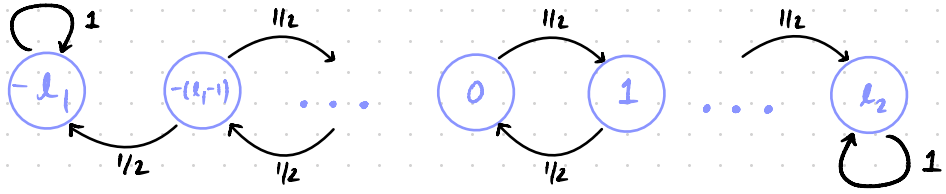
Then $P_{x,y} = \frac{1}{n}$ and $P_{y,x} = \frac{1}{n}$.

If $\pi_x = \pi_y = \frac{1}{M}$, then $\pi_x P_{x,y} = \frac{1}{Mn} = \pi_y P_{y,x}$.

This satisfies the time-reversibility condition and proves the stationary distribution is uniform.

9. Consider a sequence of independent, fair gambling games between two players. In each round a player wins a dollar with probability $\frac{1}{2}$ or loses a dollar with probability $\frac{1}{2}$. We can construct a markov chain such that the state of the system at the time t is the number of dollars won by player 1. If player 1 lost money, this number is negative. Initial state [2] is 0. It is reasonable to assume that there are numbers ℓ_1 and ℓ_2 such that player i (for $i \in \{1, 2\}$) cannot lose more that ℓ_i dollars and thus game ends when the markov chain reaches one of the two states $-\ell_1$ or ℓ_2. At this point, one of the gamblers is ruined. That is, they lost all their money. What is the probability that player 1 wins ℓ_2 dollars before losing ℓ_1 dollars.

Here, $-\ell_1$ and ℓ_2 are reccurent states for the following Markov Chain:



All other states are transient, since there is a non-zero probability of moving from each of these states to either state $-\ell_1$ or state ℓ_2.

Let P_i^t be the probability that, after t steps, the chain is at state i. For $-\ell_1 < i < \ell_2$, state i is transient and so $\lim_{t\to\infty} P_i^t = 0$.

Let q be the probability that the game ends with player 1 winning ℓ_2 dollars, so that the chain was absorbed into state ℓ_2. Then $1-q$ is the probability the chain was absorbed into state $-\ell_1$. By definition,

$$\lim_{t\to\infty} P_{\ell_2}^t = q.$$

Since each round of the gambling game is fair, the expected gain of player 1 in each step 0. Let W^t be the gain of the player 1 after t steps. Then $E(W^t) = 0$ for any t by induction.

$$\therefore \quad E(W^t) = \sum_{i=-l_1}^{l_2} i P_i^t = 0$$

and

$$\lim_{t \to \infty} E(W^t) = l_2 \cdot q + (1-q)(-l_1) = 0.$$

$$\therefore \quad q = \frac{l_1}{l_1 + l_2}.$$

10. Consider a random walk on an infinite line. At each step, the position of the particle is one of the integer points. At the next step, it moves to one of the neighbouring integral points with equal probability. Show that the expected distance of the particle from origin after n steps is $\Theta(\sqrt{n})$. (**Hint:** If $\mathbb{E}[|X|]$ is hard to analyze, one could possibly compute $\mathbb{E}[X^2]$ and then obtain $\mathbb{E}[|X|]$ using Jensen's inequality).

Let $X_i = \begin{cases} 1 & \text{if it moves to the right} \\ -1 & \text{if it moves to the left} \end{cases}$

The distance travelled by the particle X will be (after n steps)

$$X = \sum_{i=1}^{n} X_i$$

Since it moves left and right with equal probability,

$$E(X_i) = 1 \times Pr(X_i = 1) + (-1) \times Pr(X_i = -1)$$
$$= 1 \times \frac{1}{2} - 1 \times \frac{1}{2} = 0.$$

$$\therefore E(X_i) = 0.$$

Similarly, $E(X_i^2) = 1 \times \frac{1}{2} + 1 \times \frac{1}{2} = 1.$

We have,

$$X = \sum_{i=1}^{n} X_i$$

$$X^2 = \sum_{i=1}^{n} X_i^2 + 2 \sum_{1 \le i < j \le n} X_i X_j$$

$$\therefore E(X^2) = \sum_{i=1}^{n} E(X_i^2) + 2 \sum_{1 \le i < j \le n} E(X_i X_j)$$

→ linearity of Expectation.

$$\therefore E(X^2) = \sum_{i=1}^{n} 1 + 2 \sum_{1 \le i < j \le n} E(X_i)E(X_j)$$ (X_i & X_j are i.i.d.)

$$= n + 2 \cdot 0$$

$$E(X^2) = n^2$$

We have Jensen's Inequality for a convex function f as follows:

$$E(f(X)) \geq f(E(X))$$

Let $f : x \longmapsto x^2$

$$\therefore E(X^2) \geq (E(X))^2$$

$$\Rightarrow |E(X)|^2 \leq n$$

$$\Rightarrow E(|X|) \leq \sqrt{n}$$

$\therefore$ Expected distance is $\Theta(\sqrt{n})$

11. Given a set S of n elements drawn from a totally ordered universe, the median of S is an element m of S such that at least $\lfloor n/2 \rfloor$ elements in S are less than or equal to m and at least $\lfloor n/2 \rfloor + 1$ elements in S are greater than or equal to m. If the elements in S are distinct, then m is the $(\lceil n/2 \rceil)$th element in the sorted order of S. The median can be easily found deterministically in $O(n \log n)$ steps by sorting, and there is a relatively complex deterministic algorithm that computes the median in $O(n)$ time.

    The main idea of the algorithm involves sampling. The goal is to find two elements that are close together in the sorted order of S and that have the median lie between them. Specifically, we seek two elements $d, u \in S$ such that:

    - $d \leqslant m \leqslant u$ (the median is between d and u); and
    - for $C = s \in S \mid d \leqslant s \leqslant u$, $|C| = o(n/\log n)$ (the total number of elements between d and u is small).

Algorithm 1: Randomized Median Algorithm

Data: A set S of n elements over a totally ordered universe
Result: The median element of S, denoted by m

1 Pick a (multi-)set R of $\lceil n^{3/4} \rceil$ elements in S, chosen independently and uniformly at random with replacement; Sort the set R;
2 Let d be the $\left(\lfloor \frac{n^{3/4}}{2} - \sqrt{n} \rfloor\right)$th smallest element in the sorted set R;
3 Let u be the $\left(\lfloor \frac{n^{3/4}}{2} + \sqrt{n} \rfloor\right)$th smallest element in the sorted set R;
4 By comparing every element in S to d and u, compute the set $C = s \in S \mid d \leqslant s \leqslant u$ and the numbers $\ell_d = |x \in S : x < d|$ and $\ell_u = |x \in S : x > u|$;
5 If $\ell_d > n/2$ or $\ell_u > n/2$ then FAIL;
6 If $|C| \leqslant 4n^{3/4}$ then sort the set C, otherwise FAIL;
7 Output the $(\lfloor n/2 \rfloor - \ell_d + 1)$th element in the sorted order of C;

Based on this discussion, answer the following questions.

(a) Argue that the afore mentioned algorithm terminates in linear time, and if it does not output FAIL it outputs the correct median element of the input set S.

(b) Let $\mathcal{E}_1, \mathcal{E}_2$ and $\mathcal{E}_3$ be defined as follows.

$$\mathcal{E}_1 : Y_1 = |r \in R \mid r \leqslant m| < \frac{n^{3/4}}{2} - \sqrt{n}$$

$$\mathcal{E}_2 : Y_2 = |r \in R \mid r \geqslant m| < \frac{n^{3/4}}{2} - \sqrt{n}$$

$$\mathcal{E}_3 : |C| > 4n^{3/4}$$

Show that the randomized median algorithm fails if at least one of $\mathcal{E}_1, \mathcal{E}_2$, or $\mathcal{E}_3$ occurs.

(c) Using the definitions of events as above show the following.

i. $\mathbb{P}[\mathcal{E}_1], \mathbb{P}[\mathcal{E}_2] \leqslant \frac{n^{-1/4}}{4}$ and

ii. $\mathbb{P}[\mathcal{E}_3] \leqslant \frac{n^{-1/4}}{2}$.

(d) Put all the parts above together and show that the randomized median algorithm fails with a probability of at most $n^{-1/4}$.

(a) The correctness of the algorithm could give only an incorrect answer if the median were not in set C.

But then either $l_d > \frac{n}{2}$ or $l_u > n/2$ and thus step 6 of the algorithm guarantees that, in these cases, the algorithm outputs FAIL. Similarly, as long as C is sufficiently small, the total work is only linear in the size of S. Step 7 of the algorithm therefore guarantees that the algorithm does not take more than linear time; if sorting might take too long, the algorithm outputs FAIL without sorting.

(b) Failure in step 7 of the algorithm is equivalent to the event $\mathcal{E}_3$. Failure in step 6 of the algorithm occurs if and only if $l_d > n/2$ or $l_u > n/2$. But for $l_d > n/2$, the $(\frac{1}{2}n^{3/4} - \sqrt{n})$th smallest element of R must be larger than m; this is equivalent to the event $\mathcal{E}_1$. Similarly, $l_u > n/2$ is equivalent to the event $\mathcal{E}_2$.

(c) Define a random variable X_i by

$$X_i = \begin{cases} 1 & \text{if the } i^{\text{th}} \text{ sample is less than or equal to the median} \\ 0 & \text{o/w.} \end{cases}$$

The X_i are independent, since the sampling is done with replacement. Because there are $(n-1)/2 + 1$ elements in S that are less than or equal to the median, the probability that a randomly chosen element of S is less than or equal to the median

can be written as

$$Pr(X_i = 1) = \frac{(n-1)/2 + 1}{n} = \frac{1}{2} + \frac{1}{2n}.$$

The event $\mathcal{E}_1$ is equivalent to

$$Y_1 = \sum_{i=1}^{n^{3/4}} X_i < \frac{1}{2} n^{3/4} - \sqrt{n}$$

since Y_1 is a Bernoulli trials, it is a binomial RV with parameters $n^{3/4}$ and $1/2 + \frac{1}{2n}$.

$$\therefore \text{Var}(Y_1) = n^{3/4}\left(\frac{1}{2} + \frac{1}{2n}\right)\left(\frac{1}{2} - \frac{1}{2n}\right)$$

$$= \frac{1}{4} n^{3/4} - \frac{1}{4n^{5/4}}$$

$$< \frac{1}{4} n^{3/4}.$$

Applying Chebyshev's inequality then yields

$$Pr(\mathcal{E}_1) = Pr\left(Y_1 < \frac{1}{2} n^{3/4} - \sqrt{n}\right)$$

$$\leq Pr\left(|Y_1 - E(Y_1)| > \sqrt{n}\right)$$

$$< \frac{\text{Var}(Y_1)}{n}$$

$$< \frac{\frac{1}{4} n^{3/4}}{n} = \frac{1}{4} n^{-1/4}.$$

similarly, we obtain a bound for $\mathcal{E}_2$.

Now, we calculate bound for $\mathcal{E}_3$.

If $\mathcal{E}_3$ occurs, so $|C| > 4n^{3/4}$, then at least one of the following two events occurs:

$\mathcal{E}_{3,1}$: at least $2n^{3/4}$ elements of C are greater than the median

$\mathcal{E}_{2,1}$: at least $2n^{3/4}$ elements of C are smaller than the median.

We will bound the probability of the first event occurring since the second will have the same bound by symmetry.

If there are atleast $2n^{3/4}$ elements of C above the median, then the order of u in the sorted order of S was atleast $\frac{1}{2}n + 2n^{3/4}$ and thus the set R has at least $\frac{1}{2}n^{3/4} - \sqrt{n}$ samples among the $\frac{1}{2}n - 2n^{3/4}$ largest elements in S.

Let $X = \sum_{i=1}^{n^{3/4}} X_i$, where

$$X_i = \begin{cases} 1 & \text{if the } i^{th} \text{ sample is among the } \frac{1}{2}n - 2n^{3/4} \\ & \text{largest element in } S, \\ 0 & \text{o/w} \end{cases}$$

X is a binomial RV, so

$$E(X) = \frac{1}{2} n^{3/4} - 2\sqrt{n}$$

and

$$\begin{aligned} Var(X) &= n^{3/4}\left(\frac{1}{2} - 2n^{-1/4}\right)\left(\frac{1}{2} + 2n^{-1/4}\right) \\ &= \frac{1}{4} n^{3/4} - 4n^{1/4} < \frac{1}{4} n^{3/4}. \end{aligned}$$

Applying Chebyshev's inequality yields:

$$Pr(\mathcal{E}_{3,1}) = Pr\left(X \geq \frac{1}{2} n^{3/4} - \sqrt{n}\right)$$

$$\leq Pr\left(|X - E(X)| \geq \sqrt{n}\right) \leq \frac{Var(X)}{n} < \frac{1}{4} \frac{n^{3/4}}{n}$$

$$= \frac{1}{4} n^{-1/4}$$

similarly,

$$Pr(\mathcal{E}_{3,2}) \leq \frac{1}{4} n^{-1/4}.$$

$$\therefore Pr(\mathcal{E}_3) \leq Pr(\mathcal{E}_{3,1}) + Pr(\mathcal{E}_{3,2}) \leq \frac{1}{2} n^{-1/4}$$

(d) combining the bounds derived for $\mathcal{E}_1$, $\mathcal{E}_2$ & $\mathcal{E}_3$ we get

Probability that the algorithm fails

$$= Pr(\mathcal{E}_1) + Pr(\mathcal{E}_2) + Pr(\mathcal{E}_3) \leq n^{-1/4}$$