

Homework 4 Solutions

*Note: These solutions are not necessarily model answers. Rather, they are designed to be tutorial in nature, and sometimes contain a little more explanation than an ideal solution. Also, bear in mind that there may be more than one correct solution. The maximum total number of points available is 37. **Note: Problem 1 was not graded due to lack of time. Also, due to an accounting error by the reader, all scores were inadvertently reduced by 15 points. To get your correct score, add 15 points to the score written on your solutions.***

1. The algorithm is essentially the same as Algorithm 3.1 in the textbook:

Opts

1. *same as before*: Pick a set R of $n^{3/4}$ elements independently and u.a.r. from S
2. *same as before*: Sort R
3. Let d be the $(\frac{k}{n} \cdot n^{3/4} - \sqrt{n})$ th smallest element in R
4. Let u be the $(\frac{k}{n} \cdot n^{3/4} + \sqrt{n})$ th smallest element in R
5. *same as before*: By comparing every element of S to d and u , compute C, ℓ_d, ℓ_u
6. If $\ell_d > k$ or $\ell_u > n - k$ then FAIL
7. *same as before*: If $|C| \leq 4n^{3/4}$ then sort C else FAIL
8. Output the $(k - \ell_d)$ th element in C

As for the median algorithm in the book, this algorithm always runs in linear time and either outputs the correct value or FAILS. For the analysis, we just need to bound the failure probability. To do this, we will make repeated use of the following lemma:

Let Y be a binomial r.v. with parameters $n^{3/4}$ and p . Then, $E[Y] = pn^{3/4}$ and $\text{Var}[Y] = n^{3/4}p(1-p) \leq \frac{1}{4}n^{3/4}$. So, by Chebyshev's inequality, $\Pr[|Y - E[Y]| \geq \sqrt{n}] \leq \frac{1}{4}n^{-1/4}$.

Note that this same fact is used repeatedly in the analysis of the median algorithm in the textbook.

We define m to be the element of rank k in S (so m is the desired output). We will consider the following three events, which are analogous to the events with the same names in the textbook:

$$\begin{aligned} \mathcal{E}_1 : Y_1 &= |\{r \in R \mid r \leq m\}| \leq \frac{k}{n} \cdot n^{3/4} - \sqrt{n}; \\ \mathcal{E}_2 : Y_2 &= |\{r \in R \mid r \geq m\}| \leq (1 - \frac{k}{n}) \cdot n^{3/4} - \sqrt{n}; \\ \mathcal{E}_3 : |C| &> 4n^{3/4}. \end{aligned}$$

Clearly the probability that the algorithm fails is at most $\Pr[\mathcal{E}_1] + \Pr[\mathcal{E}_2] + \Pr[\mathcal{E}_3]$.

We now bound each of these probabilities in turn:

— To analyze $\Pr[\mathcal{E}_1]$, we define a r.v. X_i indicating whether the i th sample (of R) is $\leq m$. Then, $Y_1 = \sum_i X_i$ is a binomial r.v. with parameters $n^{3/4}$ and $\frac{k}{n}$. Therefore, by the above lemma,

$$\Pr[\mathcal{E}_1] = \Pr \left[Y_1 \leq \frac{k}{n} \cdot n^{3/4} - \sqrt{n} \right] < \frac{1}{4}n^{-1/4}.$$

— To analyze $\Pr[\mathcal{E}_2]$, we define a r.v. X_i indicating whether the i th sample (of R) is $\geq m$. Then, $Y_2 = \sum_i X_i$ is a binomial r.v. with parameters $n^{3/4}$ and $1 - \frac{k}{n}$. Therefore, again by the above lemma,

$$\Pr[\mathcal{E}_2] = \Pr \left[Y_2 \leq (1 - \frac{k}{n}) \cdot n^{3/4} - \sqrt{n} \right] < \frac{1}{4}n^{-1/4}.$$

— To analyze $\Pr[\mathcal{E}_3]$, as in the book we consider two events:

\mathcal{E}_3^1 : at least $2n^{3/4}$ elements of C are greater than m ;

\mathcal{E}_3^2 : at least $2n^{3/4}$ elements of C are smaller than m .

Clearly if \mathcal{E}_3 happens then so must at least one of \mathcal{E}_3^1 and \mathcal{E}_3^2 , so $\Pr[\mathcal{E}_3] \leq \Pr[\mathcal{E}_3^1] + \Pr[\mathcal{E}_3^2]$.

Let Γ_1 denote the $k - 2n^{3/4}$ smallest elements of S . \mathcal{E}_3^1 may be rewritten as: R contains $\frac{k}{n} \cdot n^{3/4} - \sqrt{n}$ elements of Γ_1 . We let X be the number of samples (of R) in Γ_1 . Then, $X = \sum_i X_i$ where X_i is a r.v. indicating whether the i th sample lies in Γ_1 . Again, X is a binomial r.v. with parameters $n^{3/4}$ and $\frac{k}{n} - 2n^{-1/4}$. Thus $E[X] = \frac{k}{n} \cdot n^{3/4} - 2\sqrt{n}$. In addition,

$$\Pr[\mathcal{E}_3^1] = \Pr\left[X \geq \frac{k}{n} \cdot n^{3/4} - \sqrt{n}\right] < \frac{1}{4}n^{-1/4}.$$

The analysis of $\Pr[\mathcal{E}_3^2]$ is symmetrical.

Putting all the above together, we see that the probability of failure is at most

$$\Pr[\mathcal{E}_1] + \Pr[\mathcal{E}_2] + \Pr[\mathcal{E}_3^1] + \Pr[\mathcal{E}_3^2] \leq n^{-1/4},$$

as required.

2. (a) By linearity of expectation, $E[X] = \sum_{i=1}^n E[X_i] = \sum_{i=1}^n \frac{1}{i} = H_n$. Since the X_i are independent, $\text{Var}[X] = \sum_{i=1}^n \text{Var}[X_i] = \sum_{i=1}^n \left(1 - \frac{1}{i}\right)$. 3pts

(b) We are interested in upper bounds for $p = \Pr[X \geq 4 \ln n]$. Note that $E[X] = H_n = \ln n + \Theta(1)$ and $\text{Var}[X] \leq H_n$. Since we are asked only for asymptotic bounds, we will omit lower order terms from our final answers. For functions $f(n), g(n)$, the notation $f(n) \sim g(n)$ means that $f(n)/g(n) \rightarrow 1$ as $n \rightarrow \infty$.

— Applying Markov's inequality, we have

3pts

$$p \leq \frac{E[X]}{4 \ln n} = \frac{H_n}{4 \ln n} \sim \frac{\ln n}{4 \ln n} = \frac{1}{4}.$$

— Applying Chebyshev's inequality, we have

3pts

$$p \leq \Pr[|X - E[X]| \geq 4 \ln n - H_n] \leq \frac{\text{Var}[X]}{(4 \ln n - E[X])^2} \leq \frac{H_n}{(4 \ln n - H_n)^2} \sim \frac{\ln n}{(3 \ln n)^2} = \frac{1}{9 \ln n}.$$

— Applying the Chernoff bound, we note that

4pts

$$p \leq \Pr[X - E[X] \geq 4 \ln n - H_n] = \Pr[X \geq (1 - \delta)\mu] \leq \exp\left(-\frac{\delta^2}{2 + \delta}\mu\right),$$

where $\mu = E[X] = H_n$ and $\delta = \frac{4 \ln n - H_n}{H_n}$. We have to be a little more careful with the asymptotics here because the expressions are in the exponent. Note that $\delta\mu = 4 \ln n - H_n \sim 3 \ln n$, and $\frac{\delta}{2 + \delta} = \frac{4 \ln n - H_n}{4 \ln n + H_n} \sim \frac{3}{5}$. Plugging these asymptotic expressions into the above bound gives

$$p \leq \exp\left(-\frac{\delta^2}{2 + \delta}\mu\right) \sim \exp\left(-\frac{9}{5} \ln n\right) = n^{-9/5}.$$

Observe how the first bound is constant, the second is inverse logarithmic, and the third is inverse polynomial (which is exponentially better than the second bound).

3. (a) Let $X_i, i = 1, 2, \dots, 10^6$ denote the casino's net loss in the i 'th game. We have

4pts

$$X_i = \begin{cases} 2 & \text{w.p. } \frac{4}{25} \\ 99 & \text{w.p. } \frac{1}{200} \\ -1 & \text{w.p. } \frac{167}{200} \end{cases} \Rightarrow e^{tX_i} = \begin{cases} e^{2t} & \text{w.p. } \frac{4}{25} \\ e^{99t} & \text{w.p. } \frac{1}{200} \\ e^{-t} & \text{w.p. } \frac{167}{200} \end{cases}$$

Therefore,

$$E[e^{tX_i}] = \frac{4}{25}e^{2t} + \frac{1}{200}e^{99t} + \frac{167}{200}e^{-t}.$$

Now, $X = X_1 + X_2 + \dots + X_{10^6}$ is the casino's net loss in the first million games, and we may compute $E[e^{tX}]$ as follows:

$$\begin{aligned} E[e^{tX}] &= E[e^{t(X_1+X_2+\dots+X_{10^6})}] \\ &= E[e^{tX_1}] \cdot E[e^{tX_2}] \cdot \dots \cdot E[e^{tX_{10^6}}] \quad \text{since the } X_i \text{ are independent} \\ &= \left(\frac{4}{25}e^{2t} + \frac{1}{200}e^{99t} + \frac{167}{200}e^{-t} \right)^{10^6}. \end{aligned}$$

(b) We are interested in the quantity

4pts

$$\begin{aligned} \Pr[X \geq 10^4] &= \Pr[e^{tX} \geq e^{10^4 t}] \\ &\leq \frac{E[e^{tX}]}{e^{10^4 t}} \quad \text{by Markov's inequality} \\ &= \left(\frac{4}{25}e^{2t} + \frac{1}{200}e^{99t} + \frac{167}{200}e^{-t} \right)^{10^6} e^{-10^4 t}. \end{aligned}$$

This bound is valid for any $t > 0$, so we are free to choose a value of t that gives the best bound (i.e., the smallest value for the expression on the right). Plugging in $t = 0.0006$ as suggested in the hint, we get the bound 0.0002. This is very small, suggesting that the casino has a problem with its machines.

Aside: It is interesting to compare the above with a direct application of Markov's inequality. To do this, we need to redefine X to be the amount of money the casino pays out, so that X is now a non-negative r.v. An easy calculation gives $E[X] = 0.98 \cdot 10^6$, and applying Markov's inequality we get the upper bound $\Pr[X \geq 10^6 + 10^4] \leq \frac{98}{101} \approx 0.97$, which is disastrously weaker than the bound we obtained from Chernoff.

4. (a) First, we process ϕ so that every variable appears at most once in each clause (eliminate repeated occurrences of a literal, and delete a clause if both a literal and its negation occur). Let n denote the number of variables, and c_i the number of variables in clause i .

3pts

- $\text{size}(x, S_i)$: return 2^{n-c_i} . The variables in clause i must be fixed to values that satisfy the clause, and the remaining variables may be assigned any value.
- $\text{select}(S_i)$: fix the variables in clause i to values that satisfy the clause; choose the values of the remaining variables independently and u.a.r.
- $\text{lowest}(x)$: for $i = 1, 2, \dots$, test if x satisfies clause i (this test is easy); return the index of the first clause that x satisfies (or "undefined" if it satisfies no clauses).

(b) The problem is that S may occupy only a tiny fraction of all possible assignments U . Thus the number

3pts

of samples t would need to be huge in order to get a good estimate of q . We give a concrete example to make this precise. Consider the very simple formula $\phi = x_1 \wedge x_2 \wedge \dots \wedge x_n$. Clearly $|S| = 1$ (the only satisfying assignment is when all n variables are TRUE). The given algorithm will output zero unless it happens to choose this assignment in one of its t samples, i.e., it outputs zero with probability $(1 - 2^{-n})^t \geq 1 - t2^{-n} \sim 1$ for any t that is only polynomial in n . Thus the relative error of the algorithm will be arbitrarily large with probability arbitrarily close to 1.

Note: It is not enough here to quote the bound from class $t = O(q/\epsilon^2 \ln(1/\delta))$, which tells us how large a sample size is sufficient to estimate the proportion q . The reason is that this is an upper bound on t , whereas here we need a lower bound. The lower bound can be derived by the very simple argument given above.

(c) Note that the first two lines of the algorithm select each pair $(x, S_i), x \in S_i$ with probability $\frac{|S_i|}{\sum_j |S_j|}$. 3pts
 $\frac{1}{|S_i|} = \frac{1}{\sum_j |S_j|}$. In other words, the first two lines pick an element u.a.r. from the *disjoint union* of the sets S_i . (Note that the goal is really to pick an element u.a.r. from the *union* $\cup_i S_i$.) Let $\Gamma = \{(x, S_i) \mid \text{lowest}(x) = i\}$. Clearly, the algorithm outputs 1 with probability $\sum_{(x, S_i) \in \Gamma} \frac{1}{\sum_j |S_j|} = \frac{|\Gamma|}{\sum_j |S_j|}$. To see that $|\Gamma| = |S|$, simply observe that every element $x \in S$ corresponds to exactly one lowest S_i , or equivalently $\Gamma = \{(x, S_{\text{lowest}(x)}) \mid x \in S\}$. It follows that the algorithm outputs 1 with probability $p = \frac{|S|}{\sum_j |S_j|}$.

(d) Clearly, for $i = 1, 2, \dots, m$ we have $|S_i| \leq |S|$. Hence, $\sum_j |S_j| \leq m|S|$, and thus $p = \frac{|S|}{\sum_j |S_j|} \geq \frac{1}{m}$. 2pts

(e) Note that X_1, \dots, X_t are independent 0-1 r.v.'s with mean p , so $E[X] = pt$ and the Chernoff bound 3pts
yields

$$\Pr[|X - pt| \geq \epsilon pt] \leq 2e^{-\epsilon^2 pt/3}.$$

The quantity on the right is bounded above by δ provided we take $t = \lceil \frac{3}{\epsilon^2 p} \ln \frac{2}{\delta} \rceil \leq \lceil \frac{3m}{\epsilon^2} \ln \frac{2}{\delta} \rceil$, using the fact from part (d) that $p \geq \frac{1}{m}$. Hence it suffices to take $t = O(\frac{m}{\epsilon^2} \log \frac{1}{\delta})$.

(f) Each iteration of the algorithm in (c) requires $O(1)$ operations, so the final algorithm takes $O(t) =$ 2pts
 $O(\frac{m}{\epsilon^2} \log \frac{1}{\delta})$ time. By definition, we have $|S| = \frac{\sum_j |S_j|}{t} \cdot tp$ and $Y = \frac{\sum_j |S_j|}{t} \cdot X$. This implies

$$Y \in [(1 - \epsilon)|S|, (1 + \epsilon)|S|] \iff X \in [(1 - \epsilon)tp, (1 + \epsilon)tp]$$

and thus

$$\Pr[Y \in [(1 - \epsilon)|S|, (1 + \epsilon)|S|]] = \Pr[X \in [(1 - \epsilon)tp, (1 + \epsilon)tp]]$$

It follows by part (e) that $\Pr[Y \in [(1 - \epsilon)|S|, (1 + \epsilon)|S|]] \geq 1 - \delta$, as required.