CMPUT 675: Randomized Algorithms **Fall 2005**

Lecture 4: Sept 20

Lecturer: Mohammad R. Salavatipour *Scribe: Bruce Fraser*

4.1 Chebyshev's Inequality (Second Moment Method)

Recall the definition of variance and standard deviation:

Definition 4.1

$$Var[X] = E[X^2] - (E[X])^2 = E[(X - E[X])^2].$$
$$\sigma(X) = \sqrt{Var[X]}.$$

We can upper bound the probability that a random variable is away from its mean using the variance.

Theorem 4.2 (Chebyshev's Inequality) *Let X be a random variable. Then for any $\lambda > 0$:*

$$\Pr[(X - E[X]) \geq \lambda] \leq \frac{Var[X]}{\lambda^2}.$$

Proof:

$$\Pr[(X - E[X]) \geq \lambda] = \Pr[(X - E[X])^2 \geq \lambda^2] \leq \frac{E[(X - E[X])^2]}{\lambda^2}$$

by Markov's Inequality. Thus:

$$\Pr[(X - E[X]) \geq \lambda] \leq \frac{Var[X]}{\lambda^2}.$$

■

Equivalently:

$$\Pr[|X - E[X]| \geq tE[X]] \leq \frac{Var[X]}{t^2(E[X])^2}.$$

The use of Markov's inequality is referred to as first moment method and the use of Chebyshev's inequality is referred to as second moment method.

Example: Coupon collector problem (revisited)

Recall the coupon collector problem from last lecture. Let X be a random variable representing the total number of trials needed to collect all n types of coupons. We computed that $E[X] = nH_n$, so using Markov's inequality $\Pr[X \geq 2E[X]] = \Pr[X \geq 2nH_n] \leq \frac{1}{2}$. We would like to use Chebyshev's inequality to get a tighter bound.

Consider a series X_i of random variables, where X_i represents the time spent with exactly i coupons. Then all X_i are independent and $X = \sum_{i=0}^{n-1} X_i$. Hence $Var[X] = Var[\sum X_i] = \sum Var[X_i]$. The X_i are geometric distributed random variables with parameters $p_i = \frac{n-i}{n}$. Therefore $E[X_i] = \frac{1}{p_i} = \frac{n}{n-i}$. It is straightforward to show that $Var[X_i] = \frac{1-p_i}{p_i^2} \leq \frac{1}{p_i^2}$. Using this we have:

$$Var[X] = \sum_{i=0}^{n-1} Var[X_i] \leq \sum_{i=0}^{n-1} \left(\frac{n}{n-i}\right)^2 = n^2 \sum_{i=0}^{n-1} \frac{1}{i^2} \leq \frac{n^2\pi^2}{6}.$$

Hence $Var[X] \in O(n^2)$.

Now we have, by Chebyshev's inequality: $\Pr[(X - E[X]) \geq nH_n] \leq \frac{O(n^2)}{n^2(\ln(n))^2}$. Hence $\Pr[(X - E[X]) \geq nH_n] \in O\left(\frac{1}{(\ln(n))^2}\right)$. Thus it is with high probability that we can collect n coupons in $O(n \ln n)$ trials.

4.2 Finding the Median in linear time

Consider the problem of finding kth smallest element of a set S containing n elements. In particular, if $k = n/2$, then we are finding the median of S. The obvious algorithm is to sort S first, which is $O(n \log n)$. But this is overkill, since we don't need to find the rank of all elements (just one particular one). We are looking for a faster, perhaps $O(n)$, algorithm for finding the median. The idea of the algorithm is simple: Sample! We hope that the median of our samples are close enough to the median of the whole set. We present the algorithm for median finding. It can be easily adjusted to find kth smallest element for any value of k.

Median finding Algorithm:

1. Sample *with replacement* $n^{\frac{3}{4}}$ elements and put in set R.
2. Sort R. (running time is $O(n^{\frac{3}{4}} \lg n)$).
3. Let $l = \frac{n^{\frac{3}{4}}}{2} - \sqrt{n}$, $h = \frac{n^{\frac{3}{4}}}{2} + \sqrt{n}$. Find lth and hth smallest elements of R, call them a and b, respectively.
4. By comparing every element of S with a and b, we find $r(a)$ (rank of a) and $r(b)$ (rank of b) in S.
5. If $r(a) > n/2$ or if $r(b) < n/2$, then Fail and stop.
6. Let $P = \{x \in S | a \leq x \leq b\}$. If $|P| > 4n^{\frac{3}{4}}$ then Fail, otherwise sort P.
7. Return the median, the $(n/2 - r(a) + 1)$th element of P.

Clearly the total running time is $2n + o(n)$. There are three ways the algorithm can fail. Can we bound the probability of each of these. Suppose m is the median. Then there are three events that can prompt failure:

1. $\mathcal{E}_1$: be the event that $a > m$
2. $\mathcal{E}_2$: be the event that $b < m$
3. $\mathcal{E}_3$: be the event that $|P| > 4n^{\frac{3}{4}}$

We will find uppr bounds on the probability of each event.

Lemma 4.3 $\Pr[\mathcal{E}_1] \leq \frac{1}{4n^{\frac{1}{4}}}$.

Proof: Define the random variable X_i:

$$X_i = \begin{cases} 1 & \text{if, for } i\text{th sample } s,\ s \le m, \\ 0 & \text{otherwise.} \end{cases}$$

Since the samplings are independent and with replacement $\Pr[X_i = 1] = 1/2$ and $\Pr[X_i = 0] = 1/2$. Consider random variable $Y_1 = \sum_{i=1}^{n^{\frac{3}{4}}} X_i$. Then event $\mathcal{E}_1$ corresponds to the event that $Y_1 < \frac{n^{\frac{3}{4}}}{2} - \sqrt{n}$. Y_1 is the sum of independent Bernoulli trials. The variance and mean of Y_1 can be computed as follows:

$$E[Y_i] = \sum^{n^{\frac{3}{4}}} E[X_i] = \frac{n^{\frac{3}{4}}}{2},$$

$$Var[Y_1] = \sigma^2[Y_1] = p(1-p)n^{\frac{3}{4}} = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right) n^{\frac{3}{4}} = \frac{n^{\frac{3}{4}}}{4}$$

Using second moment method:

$$\Pr[Y_1 < \frac{n^{\frac{3}{4}}}{2} - \sqrt{n}] \le \Pr[|Y_1 - E[Y_1]| > \sqrt{n}] \le \frac{Var[Y_1]}{n} = \frac{n^{\frac{3}{4}}}{4n} = \frac{1}{3n^{\frac{1}{4}}}$$

■

For $\mathcal{E}_2$ symmetric arguments show that $\Pr[\mathcal{E}_2] \le \frac{1}{4n^{\frac{1}{4}}}$. Finally, we bound the probability of $\mathcal{E}_3$. Obviously, if $\mathcal{E}_3$ happens then either at least $2n^{\frac{3}{4}}$ elements of P are greater than m, or at least $2n^{\frac{3}{4}}$ are smaller than m. Without loss of generality, suppose at least $2n^{\frac{3}{4}}$ of P are large than m (the arguments for the other case are symmetric). Since b is larger than all $x \in P$: $r(b) \ge \frac{n}{2} + 2n^{\frac{3}{4}}$. There are

$$n^{\frac{3}{4}} - \left(\frac{n^{\frac{3}{4}}}{2} + \sqrt{n}\right) = \frac{n^{\frac{3}{4}}}{2} - \sqrt{n}$$

elements $e \in R, e \ge b$. Consider a series of Bernoulli random variables X_i:

$$X_i = \begin{cases} 1 & \text{, if, for sample } i,\ r(i) \ge \frac{n}{2} + 2n^{\frac{3}{4}}, \\ 0 & \text{, otherwise.} \end{cases} \tag{4.1}$$

Note that

$$E[X_i] = \frac{n-q}{n} = \frac{n - (\frac{n}{2} + 2n^{\frac{3}{4}})}{n} = \frac{1}{2} - \frac{2}{n^{\frac{1}{4}}}$$

.

Now consider binomial random variable $X = \sum X_i$, representing the number of samples s such that $r(s) \ge \frac{n}{2} + 2n^{\frac{3}{4}}$. Since X is binomial:

$$E[X] = n^{\frac{3}{4}}\left(\frac{1}{2} - \frac{2}{n^{\frac{1}{4}}}\right) = \frac{n^{\frac{3}{4}}}{2} - 2\sqrt{n}$$

and

$$Var[X] = n^{\frac{3}{4}} E[X_i](1 - E[X_i]) = n^{\frac{3}{4}}\left(\frac{1}{2} - \frac{2}{n^{\frac{1}{4}}}\right)\left(\frac{1}{2} + \frac{2}{n^{\frac{1}{4}}}\right) = n^{\frac{3}{4}}\left(\frac{1}{4} - \frac{4}{n^{\frac{1}{2}}}\right) < \frac{n^{\frac{3}{4}}}{4}$$

We conclude that $\Pr[X > \frac{n^{\frac{3}{4}}}{2} - \sqrt{n}] \leq \Pr[|X - E[X]| > \frac{n^{\frac{3}{4}}}{2} - \sqrt{n}] \leq P[|X - E[X]| \geq \sqrt{n}]$.

Using Chebyshev's inequality,

$$P[|X - E[X]| \geq \sqrt{n}] \leq \frac{Var[X]}{n} < \frac{n^{\frac{3}{4}}}{4n} = \frac{1}{4n^{\frac{1}{4}}}$$

Hence $\Pr[E_3] \leq \frac{1}{2n^{\frac{1}{4}}}$.

Since for every $\mathcal{E}_i$, $\Pr[\mathcal{E}_i] \in O(n^{-\frac{1}{4}})$, we can conclude $\Pr[Failure] \in O(n^{-\frac{1}{4}})$ and that the algorithm finds the median in $O(n)$ time with high probability.