CMPUT 675: Randomized Algorithms **Fall 2005**

Lecture 7: Sept 29

Lecturer: Mohammad R. Salavatipour *Scribe: Jun Ma*

7.1 Balls and Bins

Suppose we have n bins and n balls. We throw each ball into one of the bins that is selected uniformly at random and independet of choices of other balls. The load of a bin is the number of balls in that bin after all balls have been thrown. We are interested in finding the average load and maximum load of any bin. Clearly the expected number of balls (load) in a bin is 1.

7.1.1 What is the maximum number of balls (load) in any bin?

First we compute the probability that a fix bin i contains at least k balls. Let $\mathcal{E}_i^k$ be the event that bin i contains k balls, then

$$\Pr[\mathcal{E}_i^k] = \binom{n}{k} \cdot \left(\frac{1}{n}\right)^k \cdot \left(1 - \frac{1}{n}\right)^{n-k}.$$

The first term is to choose k out of n balls to be placed in bin i, the second term is the probability that those k go to bin i, and the third term is the probability that none of the rest go to bin i. Thus probability of the the event $\mathcal{E}_i^{\geq k}$ that bin i contains at least k balls is

$$\begin{aligned}
\Pr[\mathcal{E}_i^{\geq k}] &= \sum_{j=k}^{n} \binom{n}{j} \cdot \left(\frac{1}{n}\right)^j \cdot \left(1 - \frac{1}{n}\right)^{n-j} \\
&\leq \sum_{j=k}^{n} \binom{n}{j} \cdot \left(\frac{1}{n}\right)^j \\
&\leq \sum_{j=k}^{n} \left(\frac{en}{j}\right)^j \cdot \left(\frac{1}{n}\right)^j \\
&\leq \left(\frac{e}{k}\right)^k \cdot \left(1 + \frac{e}{k} + \left(\frac{e}{k}\right)^2 + \cdots\right) \\
&\leq \left(\frac{e}{k}\right)^k \cdot \frac{1}{1 - \frac{e}{k}}.
\end{aligned}$$

Let $k = \frac{3 \ln n}{\ln \ln n}$, then

$$\begin{aligned}
\Pr[\mathcal{E}_i^{\geq k}] &\leq 2 \left(\frac{e}{\frac{3 \ln n}{\ln \ln n}}\right)^{\frac{3 \ln n}{\ln \ln n}} \\
&\leq 2e^{\frac{(1 - \ln 3 - \ln \ln n + \ln \ln \ln n) \cdot 3 \ln n}{\ln \ln n}} \\
&\leq 2e^{-3 \ln n + o(\ln n)} \\
&\leq 2e^{-2 \ln n} \\
&= \frac{2}{n^2}.
\end{aligned}$$

Therefore, using the union bound over all bins, the probability of the event $\mathcal{E}^k$ that any bin has at least k balls is at most

$$\leq \sum_{i=1}^{n} Pr\left[\mathcal{E}_i^{\geq k}\right] \leq \frac{2}{n}$$

Thus we can conclude with high probability (ore more precisely with probability at least $\left(1 - \frac{2}{n}\right)$) all bins have load of at most $\frac{3\ln n}{\ln\ln n}$. It can be shown that with high probability the maximum load of all bins is at least $\Omega\left(\frac{\ln n}{\ln\ln n}\right)$. Thus this bound is tight.

Suppose we were to use Markov's inequality to upper bound the maximum load. Then since expected load in any bin is 1, with $k = \ln n$, $\Pr\left[\mathcal{E}_i^{\geq k}\right] \leq \frac{1}{\ln n}$ which is not useful as we cannot upper bound the probability of $\mathcal{E}^k$.

If we were to use Chebychev inequality, for one ball, variance is $p(1-p) = \frac{1}{n}\left(1 - \frac{1}{n}\right)$, thus for n independent balls variance is no more than 1. Let X be the number of balls in bin i, and $t = \sqrt{n \ln n}$. Then

$$\Pr\left[|X - 1| > t\right] \leq \frac{1}{t^2}$$

and thus,

$$\Pr\left[\mathcal{E}^t\right] \leq \frac{1}{\ln n}.$$

Compare this bound for t with the one we obatined earlier.

Finally, if we were to use Chernoff bound we would have: $\Pr\left[|X - 1| > t\right] < 2^{-t}$. With $t = \log n$:

$$\Pr\left[X \geq 2\log n\right] \leq \frac{1}{n^2}$$

thus, with $t = \log n$,

$$\Pr\left[\mathcal{E}^t\right] \leq \frac{1}{\ln n}.$$

This is still weaker than the bound we obtained.

7.1.2 How many bins are empty?

First, let's compute the expected number of empty bins. Let b_i be the event that bin i is empty. Then

$$\Pr[b_i] = \left(1 - \frac{1}{n}\right)^n \approx \frac{1}{e}.$$

Define random variable X_i as:

$$X_i = \begin{cases} 1 & \text{if bin i is empty;} \\ 0 & \text{otherwise.} \end{cases}$$

Then we have:

$$\mathrm{E}\left[X\right] = \mathrm{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i}^{n} \mathrm{E}\left[X_i\right] \approx \frac{n}{e}.$$

Can we use Chernoff bound to show that w.h.p. the number of empty bins is about e^{-1}? Note that conditional on the event that bin i is empty the probability that some other bin is empty is different. So the events are not independent and we cannot apply Chernoff bound. Fortunately, there is another lemma which can be used to prove concentration of random variables around their mean when the amount of dependency of mean on each single varible is limited.

Lemma 7.1 (Azuma's Inequality) *Let X be a random variable that is determined by n trials $T_1, T_2, ..., T_i$, such that for each i and any two sequence of possible outcomes $t_1, t_2, ..., t_{i-1}, t_i$, and $t_1, t_2, ..., t_{i-1}, t'_i$,*

$$|\mathrm{E}\left[X|T_1 = t_1, ..., T_i = t_i\right] - \mathrm{E}\left[X|T_1 = t_1, ..., T_i = t'_i\right]| \leq c_i.$$

In other words changing the outcome of trial i given we have fixed the outcomes of trials $1, \ldots, i-1$, does not change the mean by more than c_i. Then

$$\Pr\left[|X - \mathrm{E}\left[X\right]| > t\right] \leq 2e^{-t^2/2\sum_{i=1}^{n} c_i^2}.$$

Azuma's inequality is one of most powerful inequalities for the tail of martingales. In the case of Balls and Bins, the n trails are throwing the n balls and the expected value of number of empty bins does not change by more than one if the outcome of one trial is changed. More formally, let T_i be the trial of throwing ball i and X be the number of empty bins. Changing the outcome of T_i does not change $\mathrm{E}\left[T^1 = t^1, ..., T^i = t^i\right]$ by more than 1. So we can apply Azuma's inequality

$$\begin{aligned} \Pr\left[|X - \mu| > \sqrt{n \ln n}\right] &\leq 2e^{\frac{-n \ln \frac{n}{2}}{\frac{2}{n}}} \\ &\leq \frac{2}{n} \end{aligned}$$

Thus with high probability the number of empty bins $X = \frac{n}{e} \pm \Theta\left(\sqrt{n \ln n}\right)$.

7.2 The Power of Two Choices

Again suppose we have n bins and n balls. Instead of choosing a single bin for each ball, two bins are picked uniformly at random (with replacement). Then the ball will be placed in the bin with lower load. We want to show that with this simple modification, the maximum load of bins drops significantly from $O(\frac{\ln n}{\ln \ln n})$ to $O(\ln \ln n)$.

Definition 7.2 *Ball i is said to have height i if it was the i'th ball to be thrown into that bin. Height of a bin is the maximum height of a ball in it.*

First let's see the intution why we expect to have a maximum load of $O(\ln \ln n)$. Denote h_i the event that a ball gets height i. Note that *always* there are at most $\frac{n}{4}$ bins with height at least 4. Thus when we throw ball i the probability that it gets height at least 5 is equal to the probability that both bins selected for that have at least 4 balls each:

$$\Pr\left[h_5\right] \leq \left(\frac{1}{4}\right)^2.$$

Therefore: E[number of bins with height $\geq 5] \leq n\left(\frac{1}{4}\right)^2$. Similarly:

$$\Pr\left[h_6\right] \leq \left(\frac{1}{4}\right)^4,$$

and in genral:

$$\Pr\left[h_{i+1}\right] \leq \left(\frac{1}{4}\right)^{2^{i-1}}.$$

For $i \approx \log\log n$, $\left(\frac{1}{4}\right)^{2^i} \approx \frac{1}{n}$. These argument are not quite correct, since each time we assume that the number of bins of height i is the same as its expected value. Below we make the formal arguments. We will use the following two lemmas in our proof. The first one follows directly from Chernoff bound. The second one is also easy to prove and is left as an exercise.

Lemma 7.3 *$Pr[B(n,p) > 2np] \le e^{\frac{-np}{3}}$.*

Lemma 7.4 *Let $x_1, x_2, ..., x_n$ be arbitrary random variables, and $Y_1, Y_2, ..., Y_n$ is such that $Y_i = Y_i(x_1, x_2, ..., x_n)$ with $Pr[Y_i = 1|x_1, ..., x_i] \le p$. Then $Pr[\sum Y_i > k] \le Pr[B(n,p) > k]$.*

Definition 7.5 *Denote by $v_i(t)$ the number of bins of height no less than i after time t (i.e. t balls have been thrown). Also let $\mu_i(t)$ be the number of balls of height no less than i after time t, and h_t be the height of ball t.*

Note that always: $v_i(t) \le \mu_i(t)$. Define event $\mathcal{E}_i$ to be the event that: $v_i(t) \le \beta_i n$, where $\beta_4 = \frac{1}{4}, \beta_{i+1} = 2\beta_i$ (for $i \ge 4$). We prove by induction that $\mathcal{E}_i$ happens w.h.p. for all $i \ge 4$.

Base case: For $i = 4$, as we mentioned earlier, always the number of bins with at least 4 balls is at most $\frac{n}{4}$. So $\mathcal{E}_4$ happens with probability 1.

Induction Step: Assume that the statement is true for all values up to i, that is $v_i(t) \le \beta_i n$ (w.h.p). We prove that for $i+1$: $v_{i+1}(t) \le \beta_{i+1} n$ (w.h.p).

Define indicator random variable Y_t as follows: $Y_t = 1$ if and only if $h_t \ge i+1$ and $v_i(t-1) \le \beta_i n$.

Note that if the second condition is true then $\Pr[Y_t = 1] \le \beta_i^2$ and if the second condition is not true then $\Pr[Y_t = 1] = 0$. Thus we can say $\Pr[Y_t = 1] \le \beta_i^2$, always. So Y_t is dominated by $B(n, \beta_i^2)$ and we can apply Lemma 7.4. Now using Lemma 7.3:

$$\Pr\left[\sum_{t=1}^{n} Y_t > 2\beta_i^2 n\right] \le e^{-\beta_i^2 n/3} = e^{-\beta_{i+1} n/6}.$$

Assume that $\beta_{i+1} n \ge 12 \ln n$. Then using the fact that for any two events A and B: $Pr(A|B) = \frac{Pr(A \bigcap B)}{Pr(B)} \le \frac{Pr(A)}{Pr(B)}$, we have:

$$\begin{aligned}\Pr[\sim \mathcal{E}_{i+1}|\mathcal{E}_i] &\le \Pr\left[\sum_{t=1}^{n} Y_t \ge \beta_{i+1} n|\mathcal{E}_i\right] \\ &\le \frac{\frac{1}{n^2}}{\Pr(\mathcal{E}_i)}\end{aligned}$$

This implies that:

$$\begin{aligned}\Pr(\sim \mathcal{E}_{i+1}) &= \Pr(\sim \mathcal{E}_{i+1}|\mathcal{E}_i)Pr(\mathcal{E}_i) + \Pr(\sim \mathcal{E}_{i+1}| \sim \mathcal{E}_i)Pr(\sim \mathcal{E}_i) \\ &\le \frac{1}{n^2} + \Pr(\sim \mathcal{E}_i)\end{aligned}$$

Since $\Pr(\sim \mathcal{E}_4) = 0$, $Pr(\sim \mathcal{E}_i)$ increases by no more than $\frac{1}{n^2}$ for every increase in i. Thus, using a simple induction $\Pr(\sim \mathcal{E}_{\log\log n}) \le \frac{\log\log n}{n^2}$. It is straightforward to check that that for $i \approx \log\log n$ steps:

$$\beta_{i+1} n < \frac{12 \ln n}{n}.$$

So we should argue what happens when $\beta_i n < \frac{12\ln n}{n}$. Let i^* be the first time whn $\beta_i < \frac{12\ln n}{n}$, Then Pr(any fixed ball enters a bin with no less than i^* balls) $< O\left(\frac{\ln^2 n}{n^2}\right)$. Similarly,

$$\Pr(\text{any two balls enter a bin with no less than } i^* \text{ balls}) \leq \binom{n}{2} O\left(\frac{\ln^4 n}{n^4}\right) \leq O\left(\frac{\ln^4 n}{n^2}\right).$$

This implies that the probability that any bin has more than $i^* + 2$ balls is at most $O(\frac{\ln^4}{n^2})$ which is $o(1)$. Thus w.h.p. the maximum load of bins is $O(\ln \ln n)$.