CMPUT 498/501: Advanced Algorithms **Fall 2025**

Lecture 8 (Sep 29, 2025): Balls and bins, power of two choices

Lecturer: Mohammad R. Salavatipour *Scribe: Kinter Ren*

8.1 Balls and Bins

We have n balls as well as n bins. For each ball, we throw it into a bin chosen uniformly at random and independent from all other balls. We try to analyze various properties such as the maximum number of balls in a bin, or number of empty bins, etc. These type of questions have applications in load balancing (scheduling) and hashing (among others).

Definition 1 *The load of a bin is the number of the balls inside it.*

We will look at some properties of bins related to their loads:

Question 1: what is the probability of any two balls landing in the same bin?

Well it's simply just $\frac{1}{n}$ because of each ball choosing bins independent uniformly at random.

Question 2: what is the expected number of empty bins?

We first consider the probability of bin i being empty for any fix i, denoted this event by b_i:

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

We define a set random variable $\{X_i\}$ as follows: let X_i equal to 1 if bin i is empty, 0 otherwise.

Let $X = \sum_i X_i$ which denotes the total number of the empty bins. We have $\mathbf{E}(X) = \sum_i \mathbf{E}(x_i) = \sum_i Pr[b_i] \approx \frac{n}{e}$.

Question 3: what is the probability of bin i having load exactly k? Or at least k?

Let $\mathcal{E}_i^k$ be the event that bin i having load exactly k and $\mathcal{E}_i^{\geq k}$ be the event that bin i having load at least k.

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

$$\begin{aligned} Pr[\mathcal{E}_i^{\geq k}] &= \textstyle\sum_{j=k}^n \binom{n}{j}(\frac{1}{n})^j(1 - \frac{1}{n})^{n-j} \\ &\leq \textstyle\sum_{j=k}^n \binom{n}{j}(\frac{1}{n})^j \\ &\leq \textstyle\sum_{j=k}^n (\frac{en}{j})^j(\frac{1}{n})^j \\ &\leq (\frac{e}{k})^k \textstyle\sum_{i=0}^{n-k} (\frac{e}{k})^i \\ &\leq (\frac{e}{k})^k(\frac{1}{1-\frac{e}{k}}) \end{aligned}$$

Question 4: what is the maximum load of all bins?

Theorem 1 *The maximum load is at most* $\frac{3 \ln n}{\ln \ln n}$ *with probability at least* $1 - \frac{2}{n}$.

Proof. Apply $k = \frac{3\ln n}{\ln\ln n}$ to the probability bound on $Pr[\mathcal{E}_i^{\geq k}]$ we just get above, then:

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

By the union bound, $Pr[\text{any bin load} \geq \frac{3\ln n}{\ln\ln n}] \leq \frac{2}{n}$.

■

Note it can be shown that the maximum load is at least $\Omega(\frac{\ln n}{\ln\ln n})$. Now we apply the concentration bounds to the maximum load.

If we use Markov's inequality: For a fixed bin, let b_i be the random variable that equals to 1 if ball i lands on that bin and 0 otherwise. Let L denote the load on that bin. We have $\mathbf{E}(L) = \sum_{i=1}^n \mathbf{E}(b_i) = \sum_{i=1}^n \frac{1}{n} = 1$. Then $Pr[L \geq k] \leq \frac{1}{k}$, which is not strong enough to bound the maximum load (cause we will apply the union bound here).

If we use Chebyshev's: We have $Var[L] = Var[\sum_{i=1}^n b_i] = \sum_{i=1}^n Var[b_i] \leq \sum_{i=1}^n \mathbf{E}(b_i) = 1$. Then $Pr[L \leq k] \leq \frac{1}{k^2}$. By taking $k = \sqrt{n\ln n}$ we have $Pr[L \leq \sqrt{n\ln n}] \leq \frac{1}{n\ln n}$. By the union bound, the probability of maximum load at most $\sqrt{n\ln n}$ is at least $1 - \frac{1}{\ln n}$. Note this is much weaker than what we have above, i.e. probability of maximum load at most $\frac{3\ln n}{\ln\ln n}$ is at least $1 - \frac{2}{n}$.

If we use Chernoff's bound: we have $Pr[|L-\mathbf{E}(L)| \geq t] < 2^{-t}$. By taking $t = 2\log n$ we have $Pr[L \geq 2\log n] \leq n$. Thus by the union bound the probability of maximum load at most $2\log n$ is at least $1 - \frac{1}{n}$. This is still weaker than $O(\frac{\ln n}{\ln\ln n})$ we obtained. Use a stonger version of Chernoff's bound we can obtain the same bound.

8.2 Azuma's inequality

Back to **Question 2**, we already show the expected number of empty bins is roughly $\frac{n}{e}$. Now the question is can we use Chernoff's bound to show it holds with high probability. The answer is no because the events $\{X_i\}$ are not independent. Imagine if bin i is empty then the probability of other bins being empty decrease because the balls are tend to land on them conditioned on bin i is empty. We will introduce a probability bound to handle the case the events are not complete independent but limited dependency then we can get a similar bound of Chernoff's.

Theorem 2 *(Azuma's) Let X be a random variable that is determined by n trails $T_1, \cdots, T_n$ such that for each i and any two sequences of possible outcomes $t_1, \cdots, , t_{i-1}, t_i$ and $t_1, \cdots, , t_{i-1}, t_i'$ the following holds: $|\mathbf{E}[X|T_1 = t_1, \cdots, T_{i=1} = t_{i=1}, T_i = t_i] - \mathbf{E}[X|T_1 = t_1, \cdots, T_{i=1} = t_{i=1}, T_i = t_i']| \leq c_i$, i.e. changing the outcome of trail i while others are fixed does not change $\mathbf{E}[X]$ by more than c_i. Then $Pr[|X - \mathbf{E}[X]| > t] \leq 2e^{-\frac{t^2}{2\sum_{i=1}^n c_i^2}}$.*

In the case of balls and bins, changing the outcome of one trail (deciding one ball landing on which bin) does not change the number of the empty bins by more than 1, thus $c_i = 1$. By directly applying Azuma's inequality we get: $Pr[|X - \frac{n}{e}| > \sqrt{n\ln n}] \leq 2e^{-\frac{n\ln n}{2n}} = \frac{2}{n}$. Thus with high probability, the number of empty bins X is in the range of $\frac{n}{e} \pm \Theta(\sqrt{n\ln n})$.

We consider another example where we can see the power of Azuma's inequality: imagine in a casino where you can bet amount up to B dollars for each round. Assume it is a fair game, i.e. the probability of win or lose is the same. If win you gain the amount bet, otherwise the bet money is gone. Overall it's like a fair coin flip with the value which equals to the amount you bet. Let X_t be the result of coin flip in round t and Y_t as the total gain after t rounds then $|Y_t - Y_{t-1}| \leq B$ and the expected value of gain in each round is zero. Applying Azuma's inequality we get $Pr[Y_t \geq \lambda] \leq 2e^{1\frac{\lambda^2}{2tB^2}}$. Note the bound holds no matter what strategy you play or how much amount money you bet in each round.

8.3 Power of two choices

For balls and bins suppose for each ball we select two bins (instead of one) uniformly at random then we place the ball in the bin with lower load. We will show this idea can reduce the maximum load significantly, i.e. to $O(\ln \ln n)$ ([2]). We introduce some notations before the analysis.

Definition 2 *Ball b has the height i if it is the i^{th} ball thrown in that bin.*

Note that with this notion the load of a bin is exactly the maximum height of balls in this bin. Now we give the intuition before we formally describe the whole analysis.

Let h_i be the event that a ball gets height i. Note the number of bins of height at least 4 is at most $\frac{n}{4}$ (otherwise the total number of balls would be larger than n). Thus the probability of ball i gets height 5 is when both bins it selects have height at least 4 balls, i.e. $Pr[h_5] \leq (\frac{1}{4})^2$. Then $\mathbf{E}$(the number of bins with height ≥ 5) $\leq \frac{n}{4^2}$. Similarly we can show $Pr[h_6] \leq (\frac{1}{4^2})^2$ and $Pr[h_i] \leq (\frac{1}{4})^{2^{i-4}}$ for general $i > 4$. Thus by taking $i \approx \log \log n$ we can see $h_i \approx \frac{1}{n}$ which is already very small probability. Note these analysis are under the assumption that the value (the number of bins with height i) are always close to their expectation. Now we give a complete analysis.

Let X be a binomial random variable sampled from $B(n,p)$, i.e. $Pr[X = i] = \binom{n}{i} p^i (1-p)^{n-i}$ for $i = 0, 1, \cdots, n$ (consider X as the number of heads flipping n biased coins with head probability p). By apply Chernoff's bound we have the following lemma:

Lemma 1 $Pr[X > 2np] \leq e^{-\frac{np}{3}}$.

Lemma 2 *Suppose $x_1, \cdots, x_n \in B(n,p)$ are binomial variables and $y_1, \cdots, y_n$ are random variable such that $y_i = y_i(x_1, \cdots, x_n)$ with $Pr[y_i = 1 | x_1, \cdots, x_n] \leq p$. Then $Pr[\sum_i y_i > k] \leq Pr[\sum_i x_i > k]$.*

This lemme shows that even if $y_1, \cdots, y_n$ are not independent random variables, as long as the conditional head probability $(y_i = 1 | x_1, \cdots, x_n)$ is still bounded by p then $\sum_i y_i$ can be upper bound by binomial variable $B(n,p)$.

Definition 3 *We define $v_i(t)$ to be the number of bins of height at least i after time t, $\mu_i(t)$ to be the number of balls of height at least i after time t and h_t to be the height of the ball thrown at time t.*

Note by the definition, $\mu_i(t) \geq v_i(t)$ holds because in a bin of height at least i there are at least one ball of height at least i.

Definition 4 *We define $\mathcal{E}_i$ to be the events that $v_i(t) \leq \beta_i n$ where $\beta_4 = \frac{1}{4}$ and $\beta_{i+1}\beta_i^2$ for $i > 4$.*

AS what we described in the intuition part, we have the following holds:

Lemma 3 *$\forall i \geq 4$, $\mathcal{E}_i$ holds with high probability.*

Proof. We prove it by induction. The base case is $i = 4$. In this case we already argue that the number of bins of height at least 4 is at most $\frac{n}{4}$, i.e. $v_4(t) \leq \frac{n}{4}$, thus $\mathcal{E}_4$ holds.

For the inductive step, assume $\mathcal{E}_i$ holds ($i \geq 4$): $v_i(t) \leq \beta_i n$. Let Y_i be random variable that equals to 1 if and only if $h_t \geq i+1$ and $v_i(t-1) \leq \beta_i n$. Notice that if $pr[h_t \geq i+1 | v_i(t-1) \leq \beta_i n] \leq \beta_i^2$. Thus $Pr[Y_t = 1] \leq \beta^2$ and Y_t is dominated by $B(n, \beta_i^2)$ as described in Lemma 2. Combined with lemma 1 we get $Pr[\sum_{t \geq 4} Y_t > 2\beta_i^2 n] \leq e^{-\frac{\beta_i^2 n}{3}} = e^{-\frac{\beta_{i+1} n}{6}}$.

Assume $\beta_{i+1} \geq \frac{12 \ln n}{n}$ for now (implying $Pr[\sum_{t \geq 4} Y_t > 2\beta_i^2 n] \leq \frac{1}{n^2}$), then using the fact that $Pr[A|B] = \frac{Pr[A \cup B]}{Pr[B]} \leq \frac{Pr[A]}{Pr[B]}$ we get $Pr[\sim \mathcal{E}_{i+1} | \mathcal{E}_i] = Pr[\sum_{t \geq 4} Y_t \geq \beta_{i+1} n | \mathcal{E}_i] \leq \frac{1}{n^2 Pr[\mathcal{E}_i]}$. Thus $Pr[\sim \mathcal{E}_{i+1}] = Pr[\sim \mathcal{E}_{i+1} | \mathcal{E}_i] Pr[\mathcal{E}_i] + Pr[\sim \mathcal{E}_{i+1} | \sim \mathcal{E}_i] Pr[\sim \mathcal{E}_i] \leq \frac{1}{n^2} + Pr[\sim \mathcal{E}_i]$. Since $Pr[\sim \mathcal{E}_4] = 0$ and $Pr[\sim \mathcal{E}_i]$ increased by $\frac{1}{n^2}$ for each step i. Up to $\beta_{i+1} \geq \frac{12 \ln n}{n}$ (which is equivalent to $i \approx \log \log n$), $\sim \mathcal{E}_i$ still happens with very small probability.

We now deal with the case $\beta_{i+1} < \frac{12 \ln n}{n}$, let i^* be the first time that $\beta_{i^*} < \frac{12 \ln n}{n}$, then notice the probability that for any fixed ball enters a bin with at least i^* balls is at most $O(\frac{\ln^2 n}{n^2})$. Then the probability that any two balls enter a bin with at least i^* balls is at most $\binom{n}{2} O(\frac{\ln^4 n}{n^4}) = O(\frac{\ln^4 n}{n^2})$, this implies $Pr[\text{any bins has more than } i^* + 2 \text{ balls }] \leq O(\frac{\ln^4 n}{n^2})$

■

8.4 Extension

We have shown a huge improvement going from choosing one bin to two bins, i.e. maximum load from $\Theta(\frac{\ln n}{\ln \ln n})$ to $\Theta(\ln \ln n)$. The natural question is: what if we choose more bins (say d bins) and place the ball in the minimum load one?

It turns out there's still improvement but nothing significant.

Theorem 3 *[1] For $d > 2$, the d choice in balls in bins problem gives the maximum load of $\frac{\ln n \ln n}{\ln d} \pm O(1)$ with probability $1 - O(\frac{1}{n})$.*

References

[1] Yossi Azar, Andrei Z Broder, Anna R Karlin, and Eli Upfal. Balanced allocations. In *Proceedings of the twenty-sixth annual ACM symposium on theory of computing*, pages 593–602, 1994.

[2] Michael Mitzenmacher. The power of two choices in randomized load balancing. *IEEE Transactions on Parallel and Distributed Systems*, 12(10):1094–1104, 2002.