CS 388R: Randomized Algorithms Fall 2017

Lecture 6 — September 19, 2017

Prof. Eric Price *Scribe: Paul Baird-Smith and Ewin Tang*

NOTE: THESE NOTES HAVE NOT BEEN EDITED OR CHECKED FOR CORRECTNESS

1 Overview

We learned from last lecture that if you put n balls in n bins i.i.d uniformly at random, then the maximum number of balls in any bin (the max load) is $\Theta(\frac{\log n}{\log \log n})$, with good concentration. Suppose we want to improve this, so for each ball, we choose two bins at random (we will assume without replacement), and place the ball in the bin with the lighter load. As we will see, this will perform much better: $O(\log \log n)$ with good concentration.

2 Intuition

Consider the number of bins that have at least k balls: call this random variable V_k. Then an immediate bound is that $V_i \leq \frac{n}{i}$, because we only have n balls to distribute. This bound is terrible, especially as i increases, but we can use the $i = 4$ case as a base case.

So how many bins can we have with at least 5 balls? Intuitively, we know that a particular ball needs to pick two bins that both have at least four balls for that ball to be placed in the fifth position or higher. So we would expect that $V_5 = n\left(\frac{V_4}{n}\right)^2 \leq \frac{1}{4^2}n$.

More generally, the trend we expect is that $V_i \leq \beta_i n$, where

$$\beta_4 = \frac{1}{4} \qquad \beta_i = \beta_{i-1}^2$$

The max load is the largest nonzero V_i, so we want to find the β_i that puts us below $\frac{1}{n}$. Notice that the β_i follow a doubly exponential relation:

$$\begin{aligned} \log\log\frac{1}{\beta_4} &= 1 \\ \log\log\frac{1}{\beta_i} &= 1 + \log\log\frac{1}{\beta_{i-1}} \\ \implies \log\log\frac{1}{\beta_i} &= i - 3 \end{aligned}$$

So the β_i we want occurs when $i \leq \log\log n + 3 = O(\log\log n)$, as expected.

3 Claim & Proof

We need to formalize our intuition above, so let us introduce finer variables. Let $V_i(t) = \#$ of bins at height at least i after t balls have been inserted.

Proposition 1. *With high probability (that is, with probability $1 - O\left(\frac{1}{n^c}\right)$, where c can be made arbitrarily large), we can bound $V_i(t)$ as follows:*

$$V_i(t) \leq \beta_i n \quad \forall t,\ \forall i \geq 4$$
$$\beta_4 = \frac{1}{4} \quad \beta_i = 2\beta_{i-1}^2$$

Even though we have added a factor of two to our β_i, using the same argument as described in the intuition, this bound allows us to make our desired statement.

Corollary 2. *With high probability, no bin has more than $O(\log \log n)$ balls.*

Now, let us prove the proposition.

Proof. We will prove by induction. For the base case $i = 4$, we know that the statement is always true, and so is true w.h.p.

Now, consider a general i.

The structure of the proof is that we will use the bound on V_i to bound V_{i+1} by cleverly choosing variables with which we can use the Chernoff bound, and then pulling back the bound on the new variables to the bound on the desired variables. This will not work for large enough i, but in these cases, bad events are so rare that a more simple bound can complete the proof.

We want to put $V_{i+1}(t)$ in terms of $V_i(t)$, and to do so, it will be helpful to consider the placement of each ball in turn.

Let h_t be the height of the tth ball inserted. Then we can say that

$$\mathbb{P}[h_t \geq i + 1 \mid \text{state at } t - 1] \geq \left(\frac{V_i(t-1)}{n}\right)^2$$

since we need to randomly choose two of the $V_i(t-1)$ bins with at least i balls in it to have a ball placed at height $i + 1$.

We want to consider the above probability only in the case that $V_i(t-1)$ obeys the desired inequality, since otherwise we cannot induct. So let Y_t be the indicator function, where it is 1 when $h_t \geq i+1$ and $V_i(t-1) \leq \beta_i n$. Then

$$\mathbb{E}[Y_t] = \mathbb{P}[Y_t = 1] \leq \mathbb{P}[h_t \geq i + 1 \mid V_i(t-1) \leq \beta_i n] \leq \left(\frac{\beta_i n}{n}\right)^2 = \beta_i^2$$

And so $\mathbb{E}[\sum Y_t] \leq n\beta_i^2$. We would like to perform Chernoff to get a high probability bound on $\sum Y_t$, since this gets us closer to our desired quantity. However, we have two issues. First, our Y_is are not independent from each other, which is a condition for the Chernoff bound we are familiar with.

Second, directly applying the Chernoff bound will not give us a bound, since we want something in terms of $\beta_i^2 \geq \mu$.

The first problem is actually not important; we have that the bound holds even if we specify the state for time $t-1$, and so this does not matter. The second problem can be resolved simply through the following line of reasoning, using the multiplicative version of the Chernoff bound:

$$\forall B \geq \mu, \mathbb{P}[\sum_t Y_t \geq 2B] \leq \mathbb{P}[\sum_t Y_t \geq (1+B/\mu)\mu] \leq e^{-\frac{(B\mu)^2}{2+B\mu}} = e^{-\frac{B}{2\mu+B}B} \leq e^{-\frac{B}{3}}$$

And so

$$\mathbb{P}[\sum Y_t \geq \beta_{i+1} n] = \mathbb{P}[\sum Y_t \geq 2\beta_i^2 n] \leq e^{-\frac{\beta_{i+1} n}{6}}$$

This can be made below any polynomial we desire by choosing an appropriate bound for β_{i+1}. For example, to make the above below $\frac{1}{n^{10}}$, we would need $\beta_{i+1} \geq 60 \log n/n$. (Note that in two places we can choose our constant, here and at the end of the proof. Since the proof is split into two cases, we should choose both our constants to be a desired c.)

We are getting closer to our desired value, but we are dealing with number of balls instead of number of bins, and we still have the conditions on other random variables. Let Q_i be the event that $V_i(t) \leq \beta_i n$ for all t. We want to bound $\mathbb{P}[\text{any } \overline{Q_i}] \leq \sum \mathbb{P}[\overline{Q_i}]$.

If Q_i occurs, then $V_i(t-1) \leq \beta_i n$, so

$$\begin{aligned}\sum Y_t &= \#\text{ of balls at positions} \geq i+1 \\ &\geq \#\text{ bins with height} \geq i+1\end{aligned}$$

Thus, we can bound the bad event as follows:

$$\begin{aligned}\mathbb{P}[\overline{Q}_{i+1} \mid Q_i] &\leq \mathbb{P}[\sum Y_t \geq \beta_i n \mid Q_i] \\ \mathbb{P}[\overline{Q}_{i+1}] &= \mathbb{P}[\overline{Q}_{i+1} \cap \overline{Q}_i] + \mathbb{P}[\overline{Q}_{i+1} \cap Q_i] \\ &\leq \mathbb{P}[\overline{Q}_i] + \mathbb{P}[\sum Y_t \geq \beta_{i+1} n] \\ &\leq \mathbb{P}[\overline{Q}_i] + e^{-\beta_{i+1} n/6} \text{ for large enough } \beta_{i+1}\end{aligned}$$

So, in the case we are dealing with, for $\beta_{i+1} \geq 60 \log n/n$, we bounded our exponential by n^{-10}, and as a result, $\mathbb{P}[\overline{Q}_{i+1}] \leq n^{-9}$.

Thus, the claim is true for $i \leq h$, where

$$\beta_{h+1} \geq \frac{60 \log n}{n} \geq \beta_{h+2}$$

and so $\beta_h \geq \sqrt{\frac{30 \log n}{n}}$. Notice that we are almost done: β_{h+1} is close to the desired value of $\frac{1}{2n}$.

To get the last portion, we simply need to notice that the number of bins we have to deal with are small enough that even choosing two randomly is very rare. Notice that the above reasoning give us that all but αn bins have height $< O(\log \log n) = h$ where $\alpha < O(\log n/n)$. Using this, we know that the chance that any ball lies at height $\geq h$ is at most α^2. So,

$$\mathbb{E}[\#\text{ of balls at height} \geq h+1] \leq n\alpha^2 \leq \frac{\log^2 n}{n}$$

This is good, but we want a high probability bound. To do so, generalize our argument as follows:

$$\mathbb{P}[\text{max height} \geq h + c] \leq \mathbb{P}[\geq c \text{ balls above height } h]$$

Finally, we again consider balls individually. Let Z_t be the event that the tth ball is at height $\geq h$. Then $\mathbb{P}[\sum Z_i \geq c] \leq \binom{n}{c} p^c \leq (\frac{enp}{c})^c$ and so

$$p = \alpha^2 \implies \mathbb{P}[\sum Z_i \geq c] \leq \left(\frac{\log^2 n}{n} * \frac{e}{c}\right)^c \leq n^{-\frac{c}{2}}$$

For any given c, we only need to add a constant number to h to get the desired bound, and so we have high probability of the bound holding in this case as well. □