COMPSCI 630: Randomized Algorithms February 12, 2018

Lecture 8

Lecturer: Rong Ge *Scribe: Patrick Terry*

1 Overview

In this lecture we will return to the balls and bins model discussed in a previous class, this time with a particular question in mind: if n balls are placed uniformly at random into n bins, what is the maximum number of balls we will find in any one bin? The lower this number, the more evenly distributed the balls are over all the bins, which is desirable for applications such as hashing functions or distribution of tasks among many machines. We will start by finding a bound for this number that exists with high probability. Then we will look at another randomized algorithm, the two choice algorithm, which leads to a more balanced distribution of the balls over all the bins, and therefore a lower number of balls in the bin with the most balls.

2 Random Distribution Using the Balls and Bins Model

First, let's think about the number of balls in a particular bin i. This number is a random variable which we'll call X_i. Note that X_i is a sum of Bernoulli Random Variables. This is because for each ball, we can define a random variable which is equal to 1 if the ball is in bin i, and 0 otherwise, and X_i is the sum of all these variables. Because of this, we can apply Chernoff's bound:

$$\Pr[X_i \geq (1+\delta)\mu] \leq (\frac{e^{\delta}}{(1+\delta)^{1+\delta}})^{\mu}$$

Now let $t = \delta + 1$. Also note that because there are n balls in n bins, there is 1 ball in each bin on average. Thus, $\mu = E[X_i] = 1$. Then the above equation simplifies to:

$$\Pr[X_i \geq t] \leq \frac{e^{t-1}}{t^t}$$

This gives us the probability that 1 particular bin will have at least t balls. To get the probability that any of the bins will have at least t balls, we can apply a union bound:

$$\Pr[\exists i, X_i \geq t] \leq n\frac{e^{t-1}}{t^t}$$

Therefore, if $n\frac{e^{t-1}}{t^t} << 1$, that means that with high probability there is no bin with at least t balls. In order for this to be true, it turns out that:

$$t \in \Theta(\frac{\log n}{\log\log n})$$

Now, in some cases this is satisfactory, but in other cases we want an even smaller t. How do we make sure the balls are more balanced?

3 Two Choice Algorithm

3.1 The Algorithm

The algorithm itself is very simple. For each ball, randomly select two bins. Put the ball in the bin that contains fewer balls.

3.2 Intuition

Why does this help us get a smaller t? First, let's get some intuition by thinking about how a bin can get at least k balls. In order for this to happen, at the time when the kth ball was added, another bin must have at least $k-1$ balls. Let α_k denote the fraction of bins with at least k balls. We know that:

$$[\#\text{ of bins with at least k balls}] \leq [\#\text{ of balls whose two bin choices both have at least k-1 balls}]$$

Therefore, $\alpha_k \leq \alpha_{k-1}^2$.
Claim: $\alpha_k \leq \frac{1}{n}$ when $k \in \Theta(\log\log n)$.

3.3 Formalizing our Intuition

We want to prove that if there are at most α_{k-1} fraction of bins with at least $k-1$ balls, then there are $\alpha_k \approx \alpha_{k-1}^2$ bins with at least k balls. To do this, let's first set up some terminology:

- Call a bin "marked" if it has at least $k-1$ balls
- Call a ball "marked" if both its bin choices are marked

We know that the number of bins with k balls must be at most the number of marked balls.

Lemma 1. *Suppose at most α_k fraction of the bins are marked. Let X be the number of balls that are marked. $\mathbb{E}[X] \leq n\alpha_{k-1}^2$ and $\Pr[X \leq 2n\alpha_{k-1}^2] \geq 1 - \frac{1}{n^3}$ when $\alpha_{k-1}^2 \geq \frac{9\log n}{n}$.*

Proof. Consider the variable $Y = Y_1 + \ldots + Y_n$, where:

$$Y = \begin{cases} 1 \text{ with probability } \alpha_{k-1}^2 \\ 0 \text{ otherwise} \end{cases}$$

By Chernoff's bound, we know that:

$$\Pr\left[Y \geq 2\mathbb{E}[Y]\right] \leq (\frac{e}{4})^{n\alpha_{k-1}^2} \leq \frac{1}{n^3} \text{ when } \alpha_{k-1}^2 \geq \frac{9\log n}{n}$$

Now, if we can show that $X \leq Y$, then we prove the same thing for X. But X and Y are random variables; what do we even mean by $X \leq Y$? There are multiple ways to define inequalities between random variables. Here are two:

- Stronger definition: $X \leq Y$ if $\Pr[X \leq Y] = 1$
- Weaker definition: $X \leq_b Y$ if for any threshold τ, $\Pr[X \leq \tau] \geq \Pr[Y \leq \tau]$

Another way to say that the weaker definition holds is to say X is stochastically dominated by Y. Because in the previous proof, we only cared about $\Pr\left[Y \geq 2\,\mathbb{E}[Y]\right]$, it is sufficient to show that this weaker definition holds.

To show that $X \leq_b Y$, let's consider an alternate way of choosing bins that is effectively doing the exact same thing as our original algorithm:

- For each ball, sort the bins in decreasing order by number of balls in each bin
- Pick 2 numbers between 1 and n uniformly at random, and select the corresponding bins
- Let $Z_i = 1$ if both numbers are smaller than $n\alpha_{k-1}$
- Let $X_i = 1$ if ball i is marked

$X_i \leq Z_i$ if there are $\leq \alpha_{k-1}$ fraction of bins marked, which means $X \leq_b Z$. Now, note that Z has the same distribution as Y. Therefore, $X \leq_b Y$. □

Now, we have proved Lemma 1, which implies that $\alpha_k \leq 2\alpha_{k-1}^2$ with high probability if $\alpha_{k-1}^2 \geq \frac{9\log n}{n}$. Remember that to prove the original claim that $\alpha_k \leq \frac{1}{n}$ when $k \in \Theta(\log\log n)$, we need to find a k such that $\alpha_k << \frac{1}{n}$. To find such a k, we will introduce and prove the following lemma:

Lemma 2. *If $\alpha_k^2 << \frac{1}{n}$, let X be the number of bins with at least $k+1$ balls.* $\Pr[X \geq 2] = O(n^2\alpha_{k-1}^4) << 1$

This lemma is obtained by explicitly computing the probability; $\Pr[X \geq 2] \leq \binom{n}{2}(\alpha_{k-1}^2)^2$. The lemma tells us that if $\alpha_{k-1}^2 << \frac{1}{n}$, with high probability there is at most 1 bin with at least k balls.

Proof. Let $\alpha_3 = \frac{1}{3}, \alpha_i = 2\alpha_{i-1}^2$.
Let ε_i be the event that at most $\alpha_i n$ bins have at least i balls.
Claim: up to $i \in \Theta(\log\log n)$, $\Pr[\varepsilon_i] \geq 1 - \frac{i}{n^3}$.
We can prove this claim by induction. When $i = 3$, $\Pr[$ at most $\frac{n}{3}$ bins have at least 3 balls$] = 1$, by averaging argument. $\Pr[\overline{\varepsilon_{i+1}}] \leq \Pr[\overline{\varepsilon_i}] + \Pr[\overline{\varepsilon_{i+1}}|\varepsilon_i]$ We know that $\Pr[\overline{\varepsilon_i}] \leq \frac{i}{n^3}$ by induction, and we know that $\Pr[\overline{\varepsilon_{i+1}}|\varepsilon_i] \leq \frac{1}{n^3}$ by our lemma. Therefore, $\Pr[\overline{\varepsilon_{i+1}}] \leq \frac{i+1}{n^3}$, meaning that $\Pr[\varepsilon_{i+1}] \geq 1 - \frac{i+1}{n^3}$, completing our proof by induction. □