CS 388R: Randomized Algorithms, Fall 2021 September 14th, 2021

Lecture 6: Power of Two Choices

Prof. Eric Price *Scribe: Dimitrios Christou, Konstantinos Stavropoulos*

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

1 Overview

In the last lecture, we studied the *Balls and Bins* problem [RS98]. We proved that if we randomly throw n-balls to n-bins, the maximum load (i.e. the maximum concentration of balls in a bin) will be $O(\frac{logn}{loglogn})$ with high probability. This was due to the fact that the expected number of bins with i balls is approximately $2^{-O(i)}$, which means that it goes down exponentially and when we get to $i = \Theta(\frac{logn}{loglogn})$, it becomes zero.

For settings like *Hash Tables* or *Load Balancing*, we want the maximum load to be as small as possible. In the *Balls and Bins* problem, we were given only **one choice**: after randomly selecting the bin, our only choice was to throw the ball there. Thus, it is natural to consider what would happen if we were given **two choices**. Consider that we pick two bins at random, and throw the ball to the bin with the least amount of balls. This strategy still involves only constant work per ball and we expect it to decrease the maximum load compared to the single choice case. But **how much**? This is the question we are going to answer in this lecture.

2 Problem Statement

We are given n-balls and n-bins for some $n \in \mathbb{N}_+$. For each ball, we pick two bins at random and throw the ball to the bin that is *lighter*, i.e. the bin with the least amount of balls. If both bins have equal amount of balls, we throw the ball to any of them. We repeat the process until all of the balls have been thrown to bins.

Question: What is the maximum number of balls in a bin (maximum load) at the end of this process?

We will first set-up some notation. We will use the term "*height*" of a bin to refer to the number of balls that are already in the bin at some time step. Then, we define:

- $v_i(t)$:= number of bins of height $\geq i$ at step t, for $t, i \in [n]$.
- h_t := height that ball t ends up at, for $t \in [n]$.

Given the state of the bins at time $t-1$ (let it be denoted by S_{t-1}), in order for the t-th ball to end up at height at least i, both the random bins chosen at time t must have height at least $i-1$

after step $t-1$. Formally:

$$\mathbb{P}[h_t \geq i | S_{t-1}] = \left(\frac{v_{i-1}(t-1)}{n}\right)^2, \tag{1}$$

since we pick a bin with height at least $i-1$ at step t with probability $\frac{v_{i-1}(t-1)}{n}$.

Intuition: The square in Equation (1) is the key factor that will reduce the expected maximum load compared to the Balls and Bins setting (in this case, we would have $\mathbb{P}[h_t \geq i | S_{t-1}] = \frac{v_{i-1}(t-1)}{n}$).

Let's assume that ϵn-bins have height at least $i-1$ at time t, that is $v_{i-1}(t) = \epsilon n$. Then, the probability that the next ball gets to height i is (by Equation (1)) ϵ^2. Therefore, we should expect approximately $\epsilon^2 n$-bins to have height at least i, as the trend established by Equation (1) suggests (because this is true for any time t). With the same argument, we would expect $\epsilon^4 n$-bins to have height at least $i+1$, $\epsilon^8 n$-bins to have height at least $i+2$ etc.

So if, for example, we had that $v_5(t) \leq \frac{1}{2}n$ (at most half the bins have height greater or equal than 5), then we would expect

$$v_{5+k}(t) \leq \left(\frac{1}{2}\right)^{2^k} n$$

and so, by setting $k \geq \log\log n$, we would get $v_{5+k}(t) \leq 1$. This shows that on expectation, the maximum load of a bin is $O(\log\log n)$, which is significantly better that the $O(\frac{\log n}{\log\log n})$ bound we proved for the Balls and Bins problem. Interestingly, we achieve this improvement by only doubling the number of work per ball.

3 Formal Proof

The intuition we built in the previous section is not strict, since we essentially argue about the final state of the system, providing an argument only about a single time step. However, the truth is not much different from our intuition. To ensure that our analysis works, we will insert some slackness to our claim, by defining the following quantities

$$\begin{cases} \beta_4 = 1/4 \\ \beta_{i+1} = 2\beta_i^2 \end{cases}$$

Intuitively, the value β_i will provide a (uniform over $t \in [n]$) upper bound for the fraction of bins that have height at least i. As we will argue shortly, β_4 is indeed such an upper bound for $i = 4$ (with probability 1). While our intuition was that the upper bound for the fraction of bins of height $i+1$ would be approximately equal to the square of the upper bound for the fraction of bins of height i, we insert a multiplicative slackness factor of 2, to get the following (high probability) claim.

Lemma 1. *For any $c > 0$, the following statement holds with probability at least $1 - n^{-c}$*

$$v_i(t) \leq \beta_i n, \text{ for any } t \in [n] \text{ and any } i \geq 4 \text{ with } \beta_i n \geq 6(c+1)\ln n.$$

Proof. We will prove our claim via induction on i.

Basis: For $i = 4$, we have that the number of bins containing 4 balls is upper bounded by $n/4$ (with probability 1), because otherwise, the number of balls should be higher than n, which is a contradiction. ✓

Step: Assume that for some $i \in [n]$ (with $\beta_i n \leq 6(1+c)\log_e n$) the following statement holds with probability at least $1 - in^{-c-1}$, for some $c > 0$.

$$v_j(t) \leq \beta_j n, \forall t \in [n], \forall j \leq i.$$

We will prove that the statement holds even if we substitute i with $i+1$, given that $\beta_{i+1} n \geq 6(c+1)\ln n$.

To this end, we need a careful manipulation, since the inductive hypothesis holds only with high probability. We will proceed with the following steps:

1. We will first provide an upper bound for the number of balls (say Y) that have height $i+1$ AND right before they are assigned to a bin, the number of variables of height i was indeed upper bounded by $\beta_i n$ (which does not hold almost surely, but only with high probability under the inductive hypothesis – however under this condition we can calculate the probabilities of interest, as we did in Equation (1)).
2. Then, we will show that Y (for which we provided the aforementioned high probability upper bound), is also with high probability a "good" (for our purposes) representative of the number of balls that have height $i+1$ *unconditionally*, due to the inductive hypothesis.

For any $t \in [n]$, we set $Y_t := \mathbb{1}\{(h_t \geq i+1) \cap (v_i(t-1) \leq \beta_i n)\}$. Then, we have that

$$\begin{aligned}\mathbb{E}[Y_t] = \Pr[Y_t = 1] = \Pr[h_t \geq i+1 | v_i(t-1) \leq \beta_i n] \cdot \Pr[v_i(t-1) \leq \beta_i n] \leq \\ \leq \Pr[h_t \geq i+1 | v_i(t-1) \leq \beta_i n] \leq \beta_i^2 \,.\end{aligned}$$

Consider, now $Y = \sum_{t=1}^{n} Y_t$. By the linearity of expectation we get: $\mathbb{E}[Y] \leq \beta_i^2 n = \beta_{i+1} n/2$.

Observe that Y_t can only be equal to 1 given that an event ($\{v_i(t-1) \leq \beta_i n\}$) that ensures an upper bound for its expectation happens. Therefore, conditioning on any realization of $Y_1, \ldots, Y_{t-1}$ would not violate the upper bound we established for the expectation of Y_t. Hence (as we have seen in a previous lecture), although $(Y_t)_t$ are not independent, we may use the Chernoff bound on Y, because we can design some family of independent random variables $(Z_t)_t$ such that $Z_t \sim Y_t$ (i.e. Z_t and Y_t are equidistributed) and $Z_t \geq Y_t$ almost surely (i.e. with probability 1); we then apply the Chernoff bound on $Z := \sum_t Z_t$, which implies a corresponding bound for Y. We call this method (of designing the family $(Z_t)_t$) *stochastic domination* (or coupling) [B88].

From the Chernoff bound, we get

$$\Pr[Y > \beta_{i+1} n \,(\geq 2\,\mathbb{E}[Y])] \leq \exp\left(-\frac{1}{3} \cdot \frac{1}{2} \beta_{i+1} n\right).$$

But we know that $\beta_{i+1} n \geq 6(c+1)\ln n$ (in fact we have selected the right hand side of this inequality to ensure that the bound that we take from this step satisfies our purposes). Therefore

$$\Pr[Y > \beta_{i+1} n] \leq n^{-c-1} \,.$$

We have proven step 1 and we proceed with step 2.

Let $Q_j := \{v_j(t) \leq \beta_j n, \forall t \in [n]\}$. By the inductive step, we have that

$$\Pr[\cap_{j \leq i} Q_j] \geq 1 - i n^{-c-1} .$$

Also, conditioned on the event that $Y \leq \beta_{i+1} n$, then Q_{i+1} cannot be false when $\cap_{j \leq i} Q_j$ is true (almost surely), because $Y_t | Q_i = \mathbb{1}\{h_t \geq i+1\}$. Therefore

$$\Pr[\neg Q_{i+1}] \leq \Pr[Y > \beta_{i+1}] + \Pr[\neg Q_i] \leq n^{-c-1} + i n^{-c-1} = (1+i) n^{-c-1} ,$$

which concludes our induction.

Then, we observe that $i \leq n$ and therefore $i n^{-c-1} \leq n^{-c}$, which concludes the proof. □

Note that by definition of β_i (which implies that $\beta_i = 2^{-2^{i-4}-1}$), if $i^\star = \Theta(\log \log n)$, then $\beta_{i^\star} n = O(\log n)$ (and Lemma 1 only works when $\beta_i n \geq \Theta(\log n)$, so we cannot go further). Therefore, we have that with high probability (say at least $1 - n^{-\alpha-1}$, $\alpha > 0$)

$$v_{i^\star}(t) \leq O(\log n), \forall t \in [n] .$$

But we have not yet shown that (w.h.p. – i.e. with high probability) no bins have height more than some value that is $O(\log \log n)$.

We have that

$$\Pr[h_t \geq i^\star] \leq \left(\frac{O(\log n)}{n} \right)^2 + n^{-\alpha-1}, \forall t \in [n]$$

We may insert the quantifier $\exists t \in [n]$ inside the probability expression by using the union bound. We get

$$\Pr[\exists t \in [n] : h_t \geq i^\star] \leq \frac{O(\log^2 n)}{n} + n^{-\alpha} ,$$

which decays almost linearly with n (if $a \geq 1$). Also, this bounds the probability that there exists some bin with height at least $i^\star$ by $O(\log^2 n / n + n^{-\alpha})$.

Providing a high probability claim for the same event is left as an ***exercise***. (**Hint:** Consider the probability that $h_t \geq i^\star + K$).

4 A Probability Puzzle

At the end of the lecture, the professor gave as a probability puzzle to solve (which is irrelevant to the rest of the lecture).

Question: Assume you are given a biased coin that succeeds with some unknown probability p. How many samples are needed in order to estimate p within a multiplicative $(1 \pm \epsilon)$ bound?

The algorithm that we are going to use for the approximation will be the *arithmetic mean* of the samples. Formally, we will draw n independent samples $X_1, \ldots, X_n$ from the coin (recall $\mathbb{P}[X_i] = p$) and output the estimation $\hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i$. Now, we need to determine how large n must be in order for the estimation to be within a multiplicative $(1 \pm \epsilon)$ bound of its true value.

Let's assume for now that we know $p \in [a, b]$ for some $0 < a < b < 1$. Fix some number of samples n and let $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the estimation. We wish to compute

$$\mathbb{P}[\hat{p} \notin [(1-\epsilon)p, (1+\epsilon)p]] = \mathbb{P}[|\hat{p} - p| > \epsilon p]$$

Observe:

- $\hat{p}$ is the sum of n independent random variables $\frac{X_i}{n}$ that are bounded in $[0, \frac{1}{n}]$.
- $\mathbb{E}[\hat{p}] = \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} X_i] = n\frac{1}{n}p = p$.

By using a multiplicative Chernoff Bound, we get:

$$\mathbb{P}[|\hat{p} - p| > \epsilon p] \leq 2e^{-\frac{\min(\epsilon, \epsilon^2)}{3}pn} = 2e^{-\frac{\epsilon^2}{3}pn}$$

since $\epsilon < 1$. So, if we want the probability of failure to be at most $\delta \in (0, 1)$, we simply need to take at least

$$n = \frac{3\log(\frac{2}{\delta})}{\epsilon^2 p}$$

samples. Since we have assumed that $p \in [a, b]$, then we know that taking $n = \frac{3\log(\frac{2}{\delta})}{\epsilon^2 a}$ samples will give the desired guarantee.

In the next lecture, we will solve the rest of the puzzle by learning how one can do the same analysis without assuming any knowledge on p and get the same sample complexity.

References

[RS98] Martin Raab, Angelika Steger. "Balls into Bins" - A Simple and Tight Analysis. *Proceedings of the Second International Workshop on Randomization and Approximation Techniques in Computer Science*, 58(1):159–170, 1998.

[B88] Vijay S. Bawa. "Research Bibliography-Stochastic Dominance: A Research Bibliography. *Manage. Sci.*, 697–712, 1988.