15-859M: Randomized Algorithms Anupam Gupta
Lec 8: Balls and Bins/Two Choices February 2, 2011

1 Balls and Bins

The setting is simple: n balls, n bins. When you consider a ball, you pick a bin independently and uniformly at random, and add the ball to that bin. In HW #2 you proved:

Theorem 1 *The max-loaded bin has* $O(\frac{\log n}{\log\log n})$ *balls with probability at least* $1 - 1/n$.

One could use a Chernoff bound to prove this, but here is a more direct calculation of this theorem: the chance that bin i has at least k balls is at most

$$\binom{n}{k}\left(\frac{1}{n}\right)^k \leq \frac{n^k}{k!}\cdot\frac{1}{n^k} \leq \frac{1}{k!} \leq 1/k^{k/2}$$

which is (say) $\leq 1/n^2$ for $k^* = \frac{8\log n}{\log\log n}$. To see this, note that

$$k^{k/2} \geq (\sqrt{\log n})^{4\log n/\log\log n} \geq 2^{2\log n} = n^2.$$

So union bounding over all the bins, the chance of some bin having more than k^* balls is $1/n$. (I've been sloppy with constants, you can do better constants by using Stirling's approximation.)

Here is a semantically identical way of looking at this calculation: let X_i be the indicator r.v. for bin i having k^* or more balls. Then $E[X_i] \leq 1/n^2$. And hence if $X = \sum_i X_i$, then $E[X] \leq 1/n$. So by Markov, $\Pr[X > 1] \leq E[X] \leq 1/n$. In other words, we again have

$$\Pr[\text{ max load is more than } \frac{8\log n}{\log\log n}] \to 0.$$

This idea of bounding the expectation of some variable X, and using that to upper bound some quantity (in this case the max-load) is said to use the *first moment method.*

1.1 Tightness of the Bound

In fact, $\Theta(\frac{\log n}{\log\log n})$ is indeed the right answer for the max-load with n balls and n bins.

Theorem 2 *The max-loaded bin has* $\Omega(\frac{\log n}{\log\log n})$ *balls with probability at least* $1 - 1/n^{1/3}$.

Here is one way to show this, via the *second moment method.* To begin, let us now lower bound the probability that bin i has at least k balls:

$$\binom{n}{k}\left(\frac{1}{n}\right)^k\left(1-\frac{1}{n}\right)^{n-k} \geq \left(\frac{n}{k}\right)^k\cdot\frac{1}{n^k}\cdot e^{-1} \geq 1/ek^k,$$

which for $k^{**} = \frac{\log n}{3 \log\log n}$ is at least $1/en^{1/3}$, since $k^k \leq (\log n)^{\log n/3\log\log n} = n^{-1/3}$. And so we expect $\Omega(n^{2/3})$ bins to have at least k^{**} balls.

Let us define some random variables: if X_i is the indicator for bin i having at least k^{**} balls, and X is the expected number of bins with at least k^{**} balls, we get that

$$E[X_i] \geq 1/en^{1/3} \quad \text{and} \quad E[X] = \Omega(n^{2/3}).$$

Alas, in general, just knowing that $E[X] \to \infty$ will not imply $\Pr[X \geq 1] \to 1$. Indeed, consider a random variable that is 0 w.p. $1 - 1/n^{1/3}$, and n otherwise—while its expectation is $n^{2/3}$, X is more and more likely to be zero as n increases. So we need some more information about X to prove our claim. And that comes from the second moment.

Let's appeal to Chebyshev's inequality:

$$\Pr[X = 0] \leq \Pr[|X - \mu| \geq \mu] \leq \frac{\mathrm{Var}(X)}{\mu^2} = \frac{\sum_i \mathrm{Var}(X_i) + \sum_{i \neq j} \mathrm{Cov}(X_i, X_j)}{E[X]^2}.$$

You have probably seen *covariance* before: $\mathrm{Cov}(Y, Z) := E[(Y - E[Y])(Z - E[Z])]$. But since the bins are negatively correlated (some bin having more balls makes it less likely for another bin to do so), the covariance $\mathrm{Cov}(X_i, X_j) \leq 0$. Moreover, since $X_i \in \{0, 1\}$, $\mathrm{Var}(X_i) \leq E[X_i] \leq 1$; by the above calculations, $E[X]^2 \geq n^{4/3}$. So summarizing, we get

$$\Pr[X = 0] \leq \frac{\sum_i \mathrm{Var}(X_i) + \sum_{i \neq j} \mathrm{Cov}(X_i, X_j)}{E[X]^2} \leq \frac{n}{E[X]^2} \leq n^{-1/3}.$$

In other words, there is a $1 - 1/n^{1/3}$ chance that some bin contains more than k^{**} balls:

$$\Pr[\text{ max load is less than } \frac{\log n}{3 \log\log n}] \to 0.$$

(Later, you will see how to use martingale arguments and Azuma-Hoeffding bounds to give guarantees on the max-load of bins. You can also use the "Poisson approximation" to show such a result, that's yet another cool technique.)

1.2 So, in Summary

If you want to show that some non-negative random variable is zero with high probability, show that it's expectation is tends to zero, and use Markov—the *first moment method.* If you want to show that it is non-zero with high probability, show that the variance divided by the squared mean tends to zero, and use Chebyshev—the *second moment method.*

1.3 Taking it to the Threshold

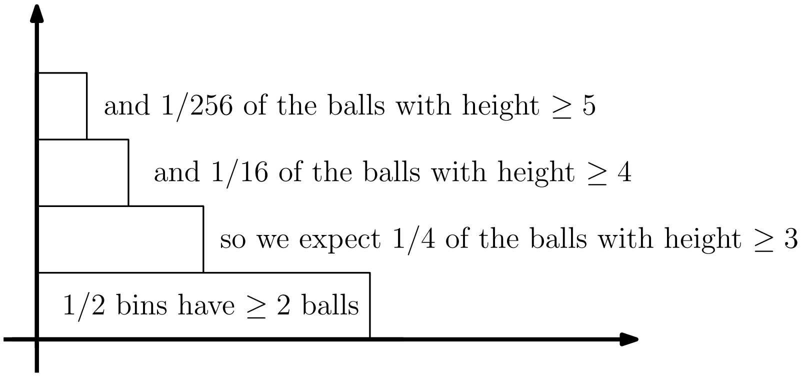
Such calculations often arise when you have a random process, and a random variable X defined in terms of a parameter k. Often you want to show that X is zero whp when k lies much below some “threshold” τ, and that X is non-zero whp when k is far above τ. The first things you should try are to see if the first and second moment methods give you rough answers. E.g., take n vertices and add each of the $\binom{n}{2}$ edges independently with probability 1/2 (also called the Erdös-Rényi graph $G(n, 1/2)$), and define X to be the expected number of cliques on k vertices. Show that $\tau = 2\log n$ is such a threshold for X.

2 The Power of Two Choices

The setting now is: n balls, n bins. However, when you consider a ball, you pick *two* bins (or in general, d bins) independently and uniformly at random, and put the ball in the less loaded of the two bins. The main theorem is:

Theorem 3 *The two-choices process gives a maximum load of* $\frac{\ln\ln n}{\ln 2} + O(1)$ *with probability at least* $1 - O(\frac{\log^2 n}{n})$.

The intuition behind the proof is the following picture:


and 1/256 of the balls with height ≥ 5
and 1/16 of the balls with height ≥ 4
so we expect 1/4 of the balls with height ≥ 3
1/2 bins have ≥ 2 balls

The actual proof is not far from this intuition. The following lemma says that if at most α fraction of the bins have at least i balls, then the fraction of bins having $i+1$ balls can indeed be upper bounded by $Bin(n, \alpha^2)$, where $Bin(n, p)$ is the Binomial random variable.

Lemma 4 *If N_i is the number of bins with load at least i, then* $\Pr[N_{i+1} > t \mid N_i \leq \alpha n] \leq \frac{\Pr[Bin(n,\alpha^2)>t]}{\Pr[N_i \leq \alpha n]}$.

Proof: For the proof, let us consider the “heights” of balls: this is the position of the ball when it comes in, if it is the first ball in its bin then its height is 1, etc. Observe that if there are t bins with load $i+1$, then there must be at least t balls with height $i+1$. I.e., if B_j is the number of balls with height at least j, then $N_j \leq B_j$, and so we'll now upper bound $\Pr[B_{i+1} > t \mid N_i \leq \alpha n] = \frac{\Pr[B_{i+1}>t \cap N_i \leq \alpha n]}{\Pr[N_i \leq \alpha n]}$.

Consider the following experiment: just before a ball comes in, an adversary is allowed to “mark” at most αn bins. Call a ball marked if both its random bins are marked. Note that when we condition on $N_i \leq \alpha n$, we know that the final number of bins with load at least i is at most αn. In this case, we can imagine the adversary marking the bins with load at least i (and maybe some more). Now the chance that a ball is marked is at least the chance that it has height $i+1$ and there are at most αn bins with height at least i. Hence, if M is the number of marked balls, we get

$$\frac{\Pr[B_{i+1} > t \cap N_i \leq \alpha n]}{\Pr[N_i \leq \alpha n]} \leq^{(*)} \frac{\Pr[M > t]}{\Pr[N_i \leq \alpha n]} = \frac{\Pr[Bin(n, \alpha^2) > t]}{\Pr[N_i \leq \alpha n]}.$$

The second equality follows from the fact that $M \sim Bin(n, \alpha^2)$. □

If you'd like to be more precise about proving (*) above, see the details in the notes from the Mitzenmacher-Upfal. (CMU/Pitt access only.)

Now we can use Chernoff to prove tail bounds on the Binomial distribution.

Lemma 5 *If* $\alpha^2 \geq 6\frac{\ln n}{n}$, *then*

$$\Pr[Bin(n, \alpha^2) > 2n\alpha^2] \leq 1/n^2.$$

Moreover, if $\alpha^2 < 6\frac{\ln n}{n}$, *then*

$$\Pr[Bin(n, \alpha^2) > 12 \ln n] \leq 1/n^2.$$

PROOF: We're interested in $X = \sum_{i=1}^n X_i$ where each $X_i = 1$ w.p. $p = \alpha^2$, and 0 otherwise. The expectation $\mu = np \geq 6 \ln n$. And the chance that this number exceeds $(1+1)\mu$ is at most

$$\exp(-\frac{\mu^2}{2\mu + \mu}) \leq \exp(-\mu/3) \leq 1/n^2,$$

which proves the first part. For the second part, $\mu < 6 \ln n$, and the probability that X exceeds $12 \ln n \geq \mu + 6 \ln n$ is at most

$$\exp(-\frac{(6 \ln n)^2}{2\mu + 6 \ln n}) \leq \exp(-2 \ln n) \leq 1/n^2,$$

as claimed. □

So, now let us define α_i to be the fraction of bins we're aiming to show have load at least i. Define $\alpha_4 = 1/4$, and $\alpha_{i+1} = 2\alpha_i^2$. (The reason it is $2\alpha_i^2$ instead of α_i^2, which is the expectation, is for some breathing room to apply Chernoff: in particular, the factor 2 comes from the first part of Lemma 5.)

For each $i \geq 4$, let $\mathcal{E}_i$ be the good event "$N_i \leq n\alpha_i$"; recall that N_i is the number of bins with load at least i. We want to lower bound the probability that this good event does not happen.

Lemma 6 *If* $\alpha_i^2 \geq 6\frac{\ln n}{n}$, *then*

$$\Pr[\neg\mathcal{E}_{i+1}] \leq i/n^2.$$

Proof: We prove this by induction. The base case is when $i = 4$, when at most $n/4$ bins can have load at least 4 (by Markov). So $\Pr[\neg\mathcal{E}_4] = 0 < 4/n^2$.

For the induction,

$$\Pr[\neg\mathcal{E}_{i+1}] \leq \Pr[\neg\mathcal{E}_{i+1} \mid \mathcal{E}_i] \Pr[\mathcal{E}_i] + \Pr[\neg\mathcal{E}_i].$$

By Lemma 4 the former term is at most $\frac{\Pr[B(n,\alpha_i^2) \geq \alpha_{i+1}]}{\Pr[\mathcal{E}_i]} \cdot \Pr[\mathcal{E}_i]$, which by Lemma 5 is at most $1/n^2$.

And by induction, $\Pr[\neg\mathcal{E}_i] \leq i/n^2$, which means $\Pr[\neg\mathcal{E}_{i+1}] \leq (i+1)/n^2$. $\square$

Consider $i^* = \min\{i \mid \alpha_i^2 < 6\frac{\ln n}{n}\}$. By the Lemma 6, $\Pr[\neg\mathcal{E}_{i^*}] \leq i^*/n^2 \leq 1/n$. Hence, with probability $1 - 1/n$, we have the number of bins with more than i^* balls in them is at most $n\alpha_{i^*}$.

We're almost done, but there's one more step to do. If this number $n\alpha_{i^*}$ were small, say $O(\log n)$, then we could have done a union bound, but this number may still be about $\Omega(\sqrt{n \log n})$. So apply Lemma 4 and the second part of Lemma 5 once more to get:

$$\begin{aligned}
\Pr[N_{i^*+1} > 12 \ln n] &\leq \Pr[N_{i^*+1} > 12 \ln n \mid \mathcal{E}_{i^*}] \Pr[\mathcal{E}_{i^*}] + \Pr[\neg\mathcal{E}_{i^*}] \\
&\leq \Pr[Bin(n, \alpha_{i^*}^2) > 12 \ln n \mid \mathcal{E}_{i^*}] \Pr[\mathcal{E}_{i^*}] + \Pr[\neg\mathcal{E}_{i^*}] \\
&\leq 1/n^2 + \Pr[\neg\mathcal{E}_{i^*}] \leq \frac{n+1}{n^2}
\end{aligned}$$

Finally, since N_{i^*+1} is so small whp, use Lemma 4 and a union bound to say that

$$\begin{aligned}
\Pr[N_{i^*+2} > 1] &\leq \Pr[B(n, \frac{(12 \ln n)^2}{n}) > 1] + \Pr[N_{i^*+1} > 12 \ln n] \\
&\leq E[B(n, \frac{(12 \ln n)^2}{n})] + \frac{n+1}{n^2} \\
&\leq O(\frac{\log^2 n}{n}).
\end{aligned}$$

Finally, the calculations in Section 2.1 show that $i^* = \frac{\ln \ln n}{\ln 2} + O(1)$, which completes the proof.

2.1 A Calculation

Since $\log_2 \alpha_4 = -2$, and $\log_2 \alpha_{i+1} = 1 + 2 \log_2 \alpha_i$, we calculate

$$\log_2 \alpha_i = -2^{i-4} - 1.$$

So, for $\log_2 \alpha_i \approx -\frac{1}{2} \log_2 n$, it suffices to set

$$i = \log_2 \log_2 n + 3 = \frac{\ln \ln n}{\ln 2} + O(1).$$

3 A Random Graphs Proof

Another way to show that the maximum load is $O(\log\log n)$—note that the constant is worse—is to use an first-priciples analysis based on properties of random graphs. We build a random graph G as follows: the n vertices of G correspond to the n bins, and the edges correspond to balls—each time we probe two bins we connect them with an edge in G. For technical reasons, we'll just consider what happens if we throw fewer balls (only $n/512$ balls) into n bins—also, let's imagine that each ball chooses two distinct bins each time.

Theorem 7 *If we throw $\frac{n}{512}$ balls into n bins using the best-of-two-bins method, the maximum load of any bin is $O(\log\log n)$ whp.*

Hence for n balls and n bins, the maximum load should be at most 512 times as much, whp. (It's as though after every $n/512$ balls, we forget about the current loads and zero out our counters—not zeroing out these counters can only give us a more evenly balanced allocation; I'll try to put in a formal proof later.)

To prove the theorem, we need two results about the random graph G obtained by throwing in $n/512$ random edges into n vertices. Both the proofs are simple but surprisingly effective counting arguments, they appear at the end.

Lemma 8 *The size of G's largest connected component is $O(\log n)$ whp.*

Lemma 9 *There exists a suitably large constant $K > 0$ such that for all subsets S of the vertex set with $|S| \geq K$, the induced graph $G[S]$ contains at most $5|S|/2$ edges, and hence has average degree at most 5, whp.*

Given the graph G, suppose we repeatedly perform the following operation in rounds:

> In each round, remove all vertices of degree ≤ 10 in the current graph.

We stop when there are no more vertices of small degree.

Lemma 10 *This process ends after $O(\log\log n)$ rounds whp, and the number of remaining vertices in each remaining component is at most K.*

Proof: Condition on the events in the two previous lemmas. Any component C of size at least K in the current graph has average degree at most 5; by Markov at least half the vertices have degree at most 10 and will be removed. So as long as we have at least K nodes in a component, we halve its size. But the size of each component was $O(\log n)$ to begin, so this takes $O(\log\log n)$ rounds before each component has size at most K. □

Lemma 11 *If a node/bin survives i rounds before it is deleted, its load due to edges that have already been deleted is at most $10i$. If a node/bin is never deleted, its load is at most $10i^* + K$, where i^* is the total number of rounds.*

PROOF: Consider the nodes removed in round 1: their degree was at most 10, so even if all those balls went to such nodes, their final load would be at most 10. Now, consider any node x that survived this round. While many edges incident to it might have been removed in this round, we claim that at most 10 of those would have contributed to x's load. Indeed, the each of the other endpoints of those edges went to bins with final load at most 10. So at most 10 of them would choose x as their less loaded bin before it is better for them to go elsewhere.

Now, suppose y is deleted in round 2: then again its load can be at most 20: ten because it survived the previous round, and 10 from its own degree in this round. OTOH, if y survives, then consider all the edges incident to y that were deleted in previous rounds. Each of them went to nodes that were deleted in rounds 1 or 2, and hence had maximum load at most 20. Thus at most 20 of these edges could contribute to y's load before it was better for them to go to the other endpoint. The same inductive argument holds for any round $i \leq i^*$.

Finally, the process ends when each component has size at most K, so the degree of any node is at most K. Even if all these edges contribute to the load of a bin, it is only $10i^* + K$. □

By Lemma 10, the number of rounds is $i^* = O(\log \log n)$ whp, so by Lemma 11 the maximum load is also $O(\log \log n)$ whp.

3.1 Missing Proofs of Lemmas

Lemma 12 *The size of G's largest connected component is $O(\log n)$ whp.*

PROOF: We have a graph with n vertices and $m = \frac{n}{512}$ edges where we connect vertices at random.

$$\begin{aligned}
\Pr[k+1 \text{ vertices connected }] &\leq \Pr[\text{ at least } k \text{ edges fall within } k+1 \text{ nodes }] \\
&\leq \binom{m}{k} \left(\frac{\binom{k+1}{2}}{\binom{n}{2}} \right)^k = \binom{m}{k} \left(\frac{k(k+1)}{n(n-1)} \right)^k \\
&\leq \binom{m}{k} \left(\frac{4k}{n} \right)^{2k}.
\end{aligned}$$

Since $k(k+1) \leq 2k^2$ and $n(n-1) \geq n^2/2$. Now the probability that any such set exists can bounded above by the union bound

$$\begin{aligned}
\Pr[\exists \text{ a connected set of size } k+1] &\leq \binom{n}{k+1} \binom{m}{k} \left(\frac{4k}{n} \right)^{2k} \\
&\leq n \left(\frac{ne}{k} \right)^k \left(\frac{ne}{512k} \right)^k \left(\frac{4k}{n} \right)^{2k} \\
&\leq n \left(\frac{e^2}{16} \right)^k \leq \frac{1}{2n} \quad \text{if } \ k = \Theta(\log n)
\end{aligned}$$

which proves the claim. □

Lemma 13 *There exists a suitably large constant $K > 0$ such that for all subsets S of the vertex set with $|S| \geq K$, the induced graph $G[S]$ contains at most $5|S|/2$ edges, and hence has average degree at most 5, whp.*

Proof:

$$\Pr[\text{ a fixed set of } k \text{ nodes gets } > \frac{5k}{2} \text{ edges }] \leq \binom{m}{5k/2}\left(\frac{4k}{n}\right)^{2\cdot 5k/2} = \binom{m}{5k/2}\left(\frac{4k}{n}\right)^{5k}.$$

By a union bound over all sets, the probability that such a set exists is

$$\begin{aligned} \Pr[\exists \text{ a bad set }] &\leq \sum_{k \geq K} \binom{n}{k}\binom{m}{5k/2}\left(\frac{4k}{n}\right)^{5k} \\ &\leq \sum_{k \geq K} \left(\frac{ne}{k}\right)^{k}\left(\frac{ne}{512(5k/2)}\right)^{5k/2}\left(\frac{k}{n}\right)^{5k} = \sum_{k \geq K}\left(\frac{k}{n}\right)^{3k/2} \alpha^k, \end{aligned}$$

where $\alpha = \frac{e^{7/2}}{80^{5/2}} < 1/2$. Now, we can break this sum into two: for small values of k, the $(k/n)^k$ term would be very small, else the α^k term would be small. Indeed, for $k \geq 2\log_2 n$, we know that $\alpha^k \leq 1/n^2$, so

$$\sum_{k=2\log n}^{n} \left(\frac{k}{n}\right)^{3k/2} \alpha^k \leq \sum_{k=2\log n}^{n} \alpha^k \leq 1/n.$$

Now for the rest:

$$\sum_{k=K}^{2\log n} \left(\frac{k}{n}\right)^{3k/2} \alpha^k \leq \sum_{k=K}^{2\log n} \left(\frac{k}{n}\right)^{3k/2} \leq 2\log n \cdot \left(\frac{2\log n}{n}\right)^{3K/2} \leq 1/n^4,$$

for $K = 3$, say. □

Bibliographic Notes: The layered induction appears in Balanced Allocations Azar, Broder, Karlin, and Upfal. The random graph analysis is in the paper Efficient PRAM Simulation on a Distributed Memory Machine by Karp, Luby, and Meyer auf der Heide; I learned it from Satish Rao. The *Always-go-left* algorithm and analysis is due to How Asymmetry Helps Load Balancing by Berthold Vöcking.

Update: Here's a survey on the various proof techniques by Mitzenmacher, Sitaraman and Richa.