Balls and Bins I : Birthday Paradox, Max Load, and Collecting Coupons[1]

- This week we are going to look at a paradigmatic model which arises as an underlying motif in many randomized algorithms : that of balls & bins. In the basic model, we have m balls which are thrown/assigned to n bins as follows : for each ball *independently* we choose one of the n bins *uniformly at random* and place it there. For $1 \leq i \leq n$, we use $\mathsf{L}_i^{(m)}$ to denote the *number* of balls that land in bin i. This is a random variable. The *load vector/profile* is the vector of random variables, $\mathsf{L}^{\vec{(m)}} := (\mathsf{L}_1^{(m)}, \ldots, \mathsf{L}_n^{(m)})$. We want to understand how the load profile "looks": there are a bunch of questions one can ask. Before we move on, observe three important things.
  - The $\mathsf{L}_i^{(m)}$'s are **identical**. This follows from symmetry of the situation.
  - The $\mathsf{L}_i^{(m)}$'s are **not independent**. After all they all sum up to m.
  - The **expected** load $\mathbf{Exp}[\mathsf{L}_i^{(m)}] = \frac{m}{n}$ for all i.
- ***The Birthday Paradox.*** This is something many of you have probably seen before[2]: in a group of around 30 individuals drawn uniformly at random, there is a $> 70\%$ chance that two of them share the same birthday. This is called the birthday "paradox" because at first glance it seems surprising : there are 365 possible birthdays (ignoring the leap-day), and so the chance a random person shares my birthday is only $\frac{1}{365}$, then how is 30 enough? The resolution of this "paradox" is of course to take a less ego-centric view : the claim is not that someone shares a birthday with me, but rather some two people share a birthday.

  The above is a balls-and-bins problem. There are n bins corresponding to the 365 birthdays. There are m balls corresponding to the 30 people. We assume everyone's birthday to be a uniform day in the year, and thus, it corresponds to the ball landing in one of the n bins u.a.r. The question is asking : what is the probability one of the bins has at least 2 balls? That is, what is $\mathbf{Pr}[\exists 1 \leq i \leq 365 \;:\; \mathsf{L}_i^{(30)} \geq 2]$?
- This calculation is elementary and not difficult. Maybe, the creativity is in coming up with the correct event definition. We are interested in the event that some bin has ≥ 2 balls. Instead, look at the *complement* event : define $\mathcal{E}$, that is *every* bin has ≤ 1 ball. We are interested in $\mathbf{Pr}[\overline{\mathcal{E}}] = 1 - \mathbf{Pr}[\mathcal{E}]$. Thus, figuring out $\mathbf{Pr}[\mathcal{E}]$ will suffice. Now comes the key definition :

  $$\mathcal{E}_i := \{\text{The } i\text{th ball lands in a bin which previously had no balls.}\}$$

  Therefore, $\mathcal{E} = \mathcal{E}_1 \wedge \mathcal{E}_2 \wedge \ldots \wedge \mathcal{E}_m$. Note: these events are **not** independent. Nevertheless, we can always write:

  $$\mathbf{Pr}[\mathcal{E}] = \mathbf{Pr}[\mathcal{E}_1] \cdot \mathbf{Pr}[\mathcal{E}_2 \mid \mathcal{E}_1] \cdot \mathbf{Pr}[\mathcal{E}_3 \mid \mathcal{E}_1 \wedge \mathcal{E}_2] \cdots \mathbf{Pr}[\mathcal{E}_m \mid \mathcal{E}_1 \wedge \mathcal{E}_2 \wedge \cdots \wedge \mathcal{E}_{m-1}] \tag{1}$$

[1] *Lecture notes by Deeparnab Chakrabarty. Last modified : 7th April, 2021*
These have not gone through scrutiny and may contain errors. If you find any, or have any other comments, please email me at deeparnab@dartmouth.edu. Highly appreciated!

[2] If not, what joy! You will see it now

Now, what is $\mathbf{Pr}\left[\mathcal{E}_t \mid \bigwedge_{i<t} \mathcal{E}_i\right]$? If the first $(t-1)$ balls have led to no collisions, they all occupy $(t-1)$ bins. Therefore, when the tth ball is being thrown, the number of *empty* bins is precisely $n-(t-1)$. Therefore,

$$\mathbf{Pr}\left[\mathcal{E}_t \mid \bigwedge_{i<t} \mathcal{E}_i\right] = \frac{n-(t-1)}{n} = 1 - \frac{t-1}{n}$$

Plugging this into (1), we get that

$$\mathbf{Pr}[\mathcal{E}] = \prod_{t=1}^{m}\left(1 - \frac{t-1}{n}\right) \tag{2}$$

- Now, if $m = 30$ and $n = 365$, then you can *exactly* calculate $\mathbf{Pr}[\mathcal{E}]$, and then $(1 - \mathbf{Pr}[\mathcal{E}])$ would exactly give you the probability that two people share the same birthday. What is more interesting is the qualitative question : if there are n bins, how big does it suffice for m to be such that we observe a collision with probability $\geq (1-\delta)$. Or in other words, $\mathbf{Pr}[\mathcal{E}] \leq \delta$?

  This can be answered using a very important inequality: $1 + z \leq e^z$ for all z. And indeed, when z is very small, this is approximately true (as the $z^2, z^3, \ldots$ are ignored in the expansion of e^z). We now apply this to (2) to get

$$\mathbf{Pr}[\mathcal{E}] = \prod_{t=1}^{m}\left(1 - \frac{t-1}{n}\right) \leq \prod_{t=1}^{m} e^{-\left(\frac{t-1}{n}\right)}$$

  The reason we took stuff to the exponent was because we had a product of a bunch of these terms. Therefore, the product is simply a sum in the exponent. And the sum is of the first $(m-1)$ natural numbers which evaluates to $\frac{m(m-1)}{2}$. Therefore, we get

$$\mathbf{Pr}[\mathcal{E}] \leq e^{-\frac{m(m-1)}{2n}}$$

  and if we want this to be $\leq \delta$, then choosing $m \approx \sqrt{2n\ln(1/\delta)}$ suffices. If we want 50% chance of a collision, then throwing $\sqrt{2\ln 2n} \approx 1.18\sqrt{n}$ many balls suffices. The important thing is the square-root. Note that in this regime the expected load on any machine is $\approx \frac{1}{\sqrt{n}} \ll 1$.

  Remark: *We also need to throw $\Omega(\sqrt{n})$ balls before we see any collision. To see this, one needs to use another analytic fact : if $z \in (0, 0.5)$, then $1 - z \geq e^{-z-z^2}$. Plug this into (2) to get a* lower *bound on $\mathbf{Pr}[\mathcal{E}]$. Using this, for how small a constant can you prove that if $m \leq c\sqrt{n}$, then $\mathbf{Pr}[\mathcal{E}] \geq 0.99$? That is, if $m \leq c\sqrt{n}$ balls are thrown, then the chances of a collision are less than 1%? A highly recommended exercise.*

- ***Maximum Load.*** The second important example in balls-and-bins comes when we are looking at the case of $m = n$. So, n balls are thrown into n bins. Just for this setting, let us use the shorthand L_i to denote $\mathsf{L}_i^{(n)}$. We expect $\mathbf{Exp}[\mathsf{L}_i] = 1$. The question is, are all loads around this expectation. Or can some loads be very large. In other words, how does $\max_i \mathsf{L}_i$ look like? The next claim is another paradigmatic application of the Chernoff bound.

Theorem 1. For large enough n, when n balls are thrown into n bins, then with probability $\geq 1 - \frac{1}{n}$, the load on *every* bin is $\leq \frac{C \ln n}{\ln \ln n}$ for some constant C.

Remark: *The constant C can be optimized, and indeed, a better constant can be obtained by a "first principles" proof. But, that is not the point of this lecture. The point is to show the dependence on n.*

Proof. Let us fix a bin i and upper bound the probability $\mathsf{L}_i \geq L$ for some parameter L. We want to show how when we set $L \approx \frac{\ln n}{\ln \ln n}$, we get the theorem. Since the L_i's are identical (but not independent) random variables, the same will be true for all i.

To evaluate L_i, let us define n indicator random variables corresponding to the n balls. We let $X_t = 1$ if the tth ball lands in the bin i thus contributing to its load. Therefore,

$$\mathsf{L}_i := \sum_{t=1}^{n} X_t$$

Note, $\mathbf{Pr}[X_t = 1] = \frac{1}{n}$ and X_t's are indeed independent. Chernoff bound (UT3) gives us (note: $\mathbf{Exp}[\mathsf{L}_i] = 1$),

$$\mathbf{Pr}[\mathsf{L}_i \geq (1+L)] \leq e^{-\frac{L \ln L}{2}} \underbrace{\leq}_{\text{want}} \delta_n \tag{3}$$

How small do we want this RHS to be? Well, for now let's call this δ_n. So, we have obtained for any i, $\mathbf{Pr}[\mathsf{L}_i \geq (1+L)] \leq \delta_n$. What is the probability that the ***maximum*** load is $\geq (1+L)$? This is where we use the simple but ubiquitous observation : the maximum is $\geq (1+L)$ if there is *some* load which is $\geq (1+L)$. And the "some" is upper bounded by the "sum" by the union bound[3]. More precisely,

$$\mathbf{Pr}[\max_i \mathsf{L}_i \geq (1+L)] = \mathbf{Pr}[\bigvee_{i=1}^{n} \{\mathsf{L}_i \geq (1+L)\}] \underbrace{\leq}_{\text{Union Bound}} \sum_{i=1}^{n} \mathbf{Pr}[\mathsf{L}_i \geq (1+L)] \underbrace{\leq}_{(3)} n\delta_n$$

Now we know how small δ_n needs to be. It needs to be such that $n\delta_n \leq \frac{1}{n}$. That would give the theorem. That is, $\delta_n \leq \frac{1}{n^2}$. Plugging this back into (3), we see that we need

$$e^{-\frac{L \ln L}{2}} \leq \frac{1}{n^2} \underbrace{\Rightarrow}_{\text{taking natual log and manipulating}} L \ln L \geq 4 \ln n$$

So, for how small an L do we have $L \ln L \geq 4 \ln n$? Ignore the 4 for now. Then clearly $L = \ln n$ would suffice; but for this the LHS would have an extra *multiplicative* $\ln \ln n$. And this is the reason why the correct answer is of the order $L = \frac{\ln n}{\ln \ln n}$; the denominator corrects for the $\ln L$ term.

Claim 1. For large enough n, if $L = \frac{8 \ln n}{\ln \ln n}$, then $L \ln L \geq 4 \ln n$.

Proof. $\ln L = \ln(C \ln n) - \ln(\ln \ln n)$. When n is large enough[4], we have $\ln \ln n \geq \frac{\ln \ln \ln n}{2}$. Thus, for large enough n, we have $\ln L \geq \frac{\ln \ln n}{2}$, implying $L \ln L \geq 4 \ln n$. □

[3] $\mathbf{Pr}[A \vee B] \leq \mathbf{Pr}[A] + \mathbf{Pr}[B]$

[4] $n \geq e^{e^e}$ suffices

This completes the proof of the theorem with $C = 8$. Once again, the constants are not the best, and once again, that is not the point. $\square$

In a later lecture, we will prove that this $\frac{\ln n}{\ln \ln n}$ is not only an upper bound but a lower bound as well. That is, whp the maximum load is also $\geq \frac{C' \ln n}{\ln \ln n}$ for some other constant C'. The qualitative message is important : although we expect every bin to have 1 ball, there will, with high probability, some bin with $\approx \frac{\ln n}{\ln \ln n}$ balls. But the max load is no higher (which we saw above).

- ***The Coupon Collector Problem.*** The third example is a kind of a "flip process". Imagine we are throwing balls and stop only when *all* bins have at least one ball. How many balls do we need to throw? Or in other words, how large does m need to be such that $\mathsf{L}_i^{(m)} \geq 1$ for *every* $1 \leq i \leq n$, with probability say $\geq 50\%$?

  Once again, when $m = n$, we expect the load of every bin to be 1. A gut instinct might be to say when $m = 2n$ or cn for some constant c, we would get a ball in each bin with probability 50%. This is wrong. The reason is this : as the bins get filled up, the chance that the next ball fills an empty bin reduces. And thus, it takes much longer than n time to fill up all the bins.

- Let us first do a slick and *exact* calculation of the *expected time* to fill all the bins. This analysis, akin to Karp's analysis of QUICKSORT, is something that anyone taking a randomized algorithms course should just know. So, this is perhaps a obligatory detour we must do. But it will be worth it. Once again, the key insight is in the definitions.

Theorem 2. The expected number of balls that needs to be thrown before every one of the n bins has at least one ball is precisely nH_n, where $H_n = 1 + \frac{1}{2} + \cdots + \frac{1}{n}$ is the nth Harmonic number.

Proof. Let Z be the random number of balls that need to be thrown before all the bins obtain one ball. We are going to write Z as a sum of a bunch of random variables. Let $\mathcal{E}_i$ be the event exactly i bins have at least one ball. Let Z_i denote the number of balls thrown *between* $\mathcal{E}_{i-1}$ and $\mathcal{E}_i$. That is, Z_i is the number of balls that were thrown to make the number of filled bins go up from $(i-1)$ to i. So, $Z_1 = 1$ (the first ball is always going to be in an erstwhile empty bin). $Z_2 = 1$ if the second ball is in the empty bin, but there is an $\frac{1}{n}$ chance that $Z_2 > 1$. Note that

$$Z = \sum_{i=1}^{n} Z_i \quad \Rightarrow \quad \mathbf{Exp}[Z] = \sum_{i=1}^{n} \mathbf{Exp}[Z_i] \tag{4}$$

What is $\mathbf{Exp}[Z_i]$? Well, how does the variable Z_i look like? What is the probability $Z_i = 1$? For this to occur, right after the $(i-1)$th bin is filled, the next ball lands in an empty bin. The number of empty bins at that time is $n - (i-1)$. Therefore, the probability of that is $p_i = \frac{n-(i-1)}{n}$. So, $\mathbf{Pr}[Z_i = 1] = p_i$.

What is the probability $Z_i = 2$. Well, the first ball after $\mathcal{E}_{i-1}$ missed an empty bin, and this occurs with probability $(1 - p_i)$. But the next ball does get to an empty bin. This probability, however, is *again* p_i. Thus, $\mathbf{Pr}[Z_i = 2] = (1 - p_i)p_i$. And now you can see that the Z_i is a ***geometric random variable*** with parameter p_i. And thus,

$$\mathbf{Exp}[Z_i] = \frac{1}{p_i} = \frac{n}{n-(i-1)} \underbrace{\Rightarrow}_{(4)} \mathbf{Exp}[Z] = \sum_{i=1}^{n} \frac{n}{n-i+1} = nH_n$$ $\square$