Randomized Algorithms

Lecture 9: the birthday paradox, and balls in bins

Kousha Etessami

The Birthday Problem

Birthday problem

There are 30 people in a room. I am willing to bet you that "at least two people in the room have the same birthday".

Should you take my bet? (I offer even odds.)

The Birthday Problem

Birthday problem

There are 30 people in a room. I am willing to bet you that "at least two people in the room have the same birthday".

Should you take my bet? (I offer even odds.)

In order words, you have to calculate:
is there at least $1/2$ probability that no two people will have the same birthday in a room with 30 people?
(We are implicitly assuming that these people's birthdays are independent and uniformly distributed throughout the 365(+1) days of the year, taking into account leap years.)

Toward a solution to the Birthday problem:

Question: What is the probability, p_m, that m people in a room all have different birthdays?

Toward a solution to the Birthday problem:

Question: What is the probability, p_m, that m people in a room all have different birthdays?

We can equate the birthdays of m people to a list $(b_1, \ldots, b_m)$, with each $b_i \in \{1, \ldots, 366\}$.

We are assuming each list in $B = \{1, \ldots, 366\}^m$ is equally likely.

Note that $|B| = 366^m$. What is the size of

$$A = \{(b_1, \ldots, b_m) \in B \mid b_i \neq b_j \text{ for all } i \neq j, i, j \in \{1, \ldots, m\}\} \ ?$$

This is simply the # of m-permutations from a set of size 366.

Thus $\quad |A| = 366 \cdot (366 - 1) \ldots (366 - (m - 1))$.

Thus, $p_m = \frac{|A|}{|B|} = \prod_{i=1}^{m} \frac{366-i+1}{366} = \prod_{i=1}^{m}(1 - \frac{i-1}{366})$.

By brute-force calculation, $p_{30} = 0.2947$. Thus, the probability that at least two people do have the same birthday in a room with 30 people is $1 - p_{30} = 0.7053$.

So, you shouldn't have taken my bet! Not even for 23 people in a room, because $1 - p_{23} = 0.5063$. But $1 - p_{22} = 0.4745$.

A general result underlying the birthday paradox: Balls in Bins

Theorem: Suppose that each of $m \geq 1$ balls is independently and uniformly at random placed in one of $n \geq 1$ bins. If

$$m \geq (1.1775 \cdot \sqrt{n}) + 1$$

then the probability that two balls go into the same bin is greater than 1/2.

Proof:

The probability that m balls all go in different bins, when the bin for each ball is chosen independently and u.a.r., from among n bins, is:

$$\prod_{i=1}^{m-1}(1-\frac{i}{n}) \quad \leq \quad \prod_{i=1}^{m-1} e^{-(i/n)} = e^{-\frac{1}{n}\sum_{i=1}^{m-1} i} \; = \; e^{-\frac{m(m-1)}{2n}}$$

So we want m to be big enough so that $e^{-\frac{m(m-1)}{2n}} < 1/2$.

Taking logs, and negating, this is equivalent to

$$\frac{m(m-1)}{2n} > \ln 2 \quad \Longleftrightarrow \quad m(m-1) > (2\cdot\ln 2)\cdot n$$

Thus, since $m(m-1) > (m-1)^2$, it suffices if

$$(m-1)^2 \geq (2\cdot\ln 2)\cdot n \quad \Longleftrightarrow \quad (m-1) \geq \sqrt{(2\cdot\ln 2)}\cdot\sqrt{n}$$

Thus, since $\sqrt{(2\ln 2)} = 1.177410\ldots \leq 1.1775$, it suffices if:

$$m \geq (1.1775\cdot\sqrt{n}) + 1. \quad \square$$

Note that this implies that:

- when there are $n = 366$ bins (possible birthdays),
- if there were at least $m = 1.1775 \cdot \sqrt{366} + 1 = 23.5269$ balls (people), then we have probability $\geq 1/2$ that two balls (two people) share a bin (share their birthday).
  This is not quite as good as the bound we obtained for 366 by exhaustive calculation, which showed 23 people suffice to have probability $\geq 1/2$, of two people with the same birthday, but it is close. (The bound in the proof of the theorem is a bit loose, because for simplicity we used the inequality $m(m-1) > (m-1)^2$.)

Balls into Bins

- m balls, n bins, and balls thrown uniformly at random and independently into bins (usually one at a time).
- The bins have no upper limit on capacity.
- Can be viewed as a (uniformly) random function, $f : [m] \longrightarrow [n]$.
- Common model of random assignment/allocation, and their effects on overall *load* and *load balance*.
- Also crucial for analysis of hashing and (idealized) hash functions.

Balls into Bins

- m balls, n bins, and balls thrown uniformly at random and independently into bins (usually one at a time).
- The bins have no upper limit on capacity.
- Can be viewed as a (uniformly) random function, $f : [m] \to [n]$.
- Common model of random assignment/allocation, and their effects on overall *load* and *load balance*.
- Also crucial for analysis of hashing and (idealized) hash functions.

Many related questions:

- How many balls do we need (in expectation) to cover all bins?
  (Coupon collector, *surjective mapping*)
- How many balls will lead (with probability $> 1/2$) to a collision?
  (Birthday paradox, *injective mapping*)
- What is the (expected) maximum load of any bin?
  (Load balancing)

Balls into Bins: maximum load

Goal: bound the maximum load of the "Balls into Bins" model in the case when $m = n$. For any bin $i \in [n]$, its load, denoted X_i, has expectation

$$\mathrm{E}[X_i] = \sum_{j=1}^{n} \mathrm{E}[X_{ij}] = n \cdot \frac{1}{n} = 1.$$

Balls into Bins: maximum load

Goal: bound the maximum load of the "Balls into Bins" model in the case when $m = n$. For any bin $i \in [n]$, its load, denoted X_i, has expectation

$$\mathrm{E}[X_i] = \sum_{j=1}^{n} \mathrm{E}[X_{ij}] = n \cdot \frac{1}{n} = 1.$$

Let $X_i > T$ be our "bad events" for some threshold T. Then to show that whp everyone's load is $\leq T$, via the union bound, we need to at least upper bound the bad event like this

$$\Pr[X_i > T] \leq \frac{1}{n^2}.$$

Markov's inequality gives $\Pr[X_i > T] \leq \frac{1}{T}$, but is not good enough.
Nor is Chebyshev.

Balls into Bins: maximum load

Goal: bound the maximum load of the "Balls into Bins" model in the case when $m = n$. For any bin $i \in [n]$, its load, denoted X_i, has expectation

$$\mathrm{E}[X_i] = \sum_{j=1}^{n} \mathrm{E}[X_{ij}] = n \cdot \frac{1}{n} = 1.$$

Let $X_i > T$ be our "bad events" for some threshold T. Then to show that whp everyone's load is $\leq T$, via the union bound, we need to at least upper bound the bad event like this

$$\Pr[X_i > T] \leq \frac{1}{n^2}.$$

Markov's inequality gives $\Pr[X_i > T] \leq \frac{1}{T}$, but is not good enough.
Nor is Chebyshev.

Suitable Chernoff bounds for "negatively correlated" r.v.'s can be made to work here, since X_i's are "negatively correlated", but we didn't state such Chernoff bounds.
Instead, we will do a quicker "ad hoc" proof for the upper bound.

Balls into Bins maximum load

Lemma (5.1)

Let n balls be thrown independently and uniformly at random *into n bins. Then for sufficiently large n, the maximum load is bounded above by* $\frac{3\ln(n)}{\ln\ln(n)}$ *with probability at least* $1 - \frac{1}{n}$.

[1]Stirling: $\sqrt{2\pi n}\left(\frac{n}{e}\right)^n \cdot e^{1/(12n+1)} \leq n! \leq \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \cdot e^{1/(12n)}$.

Balls into Bins maximum load

Lemma (5.1)

Let n balls be thrown independently and uniformly at random *into n bins. Then for sufficiently large n, the maximum load is bounded above by $\frac{3\ln(n)}{\ln\ln(n)}$ with probability at least $1 - \frac{1}{n}$.*

Proof: The probability that bin i receives $\geq M$ balls is *at most*

$$\binom{n}{M}\left(\frac{1}{n}\right)^M.$$

[1] Stirling: $\sqrt{2\pi n}\left(\frac{n}{e}\right)^n \cdot e^{1/(12n+1)} \leq n! \leq \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \cdot e^{1/(12n)}$.

Balls into Bins maximum load

Lemma (5.1)

Let n balls be thrown independently and uniformly at random *into n bins. Then for sufficiently large n, the maximum load is bounded above by* $\frac{3\ln(n)}{\ln\ln(n)}$ *with probability at least* $1 - \frac{1}{n}$.

Proof: The probability that bin i receives $\geq M$ balls is *at most*

$$\binom{n}{M}\left(\frac{1}{n}\right)^M .$$

But $\binom{n}{M} = \frac{n!}{M!(n-M)!}$ satisfies (e.g., using Stirling's approximation[1] of $n!$)

$$\left(\frac{n}{M}\right)^M \leq \binom{n}{M} \leq \frac{n^M}{M!} \leq \left(\frac{en}{M}\right)^M .$$

Hence, bin i gets $\geq M$ balls with probability at most

$$\binom{n}{M}\left(\frac{1}{n}\right)^M \leq \left(\frac{en}{nM}\right)^M = \left(\frac{e}{M}\right)^M .$$

[1] Stirling: $\sqrt{2\pi n}\left(\frac{n}{e}\right)^n \cdot e^{1/(12n+1)} \leq n! \leq \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \cdot e^{1/(12n)}$.

Balls into Bins maximum load

Proof of Lemma 5.1 cont'd.

Bin i gets $\geq M$ bins with probability at most $\left(\frac{e}{M}\right)^M$.

Balls into Bins maximum load

Proof of Lemma 5.1 cont'd.

Bin i gets $\geq M$ bins with probability at most $\left(\frac{e}{M}\right)^M$.

Let $M := \frac{3\ln(n)}{\ln\ln(n)}$. Then the probability that *any* bin gets $\geq M$ balls is (using the Union bound) at most

$$n \cdot \left(\frac{e \cdot \ln\ln(n)}{3\ln(n)}\right)^{\frac{3\ln(n)}{\ln\ln(n)}} \leq n \cdot \left(\frac{\ln\ln(n)}{\ln(n)}\right)^{\frac{3\ln(n)}{\ln\ln(n)}} = e^{\ln(n)} \left(\frac{\ln\ln(n)}{\ln(n)}\right)^{\frac{3\ln(n)}{\ln\ln(n)}}.$$

Balls into Bins maximum load

Proof of Lemma 5.1 cont'd.

Bin i gets $\geq M$ bins with probability at most $\left(\frac{e}{M}\right)^M$.

Let $M := \frac{3\ln(n)}{\ln\ln(n)}$. Then the probability that *any* bin gets $\geq M$ balls is (using the Union bound) at most

$$n \cdot \left(\frac{e \cdot \ln\ln(n)}{3\ln(n)}\right)^{\frac{3\ln(n)}{\ln\ln(n)}} \leq n \cdot \left(\frac{\ln\ln(n)}{\ln(n)}\right)^{\frac{3\ln(n)}{\ln\ln(n)}} = e^{\ln(n)} \left(\frac{\ln\ln(n)}{\ln(n)}\right)^{\frac{3\ln(n)}{\ln\ln(n)}} .$$

We can rewrite this as

$$\begin{aligned} e^{\ln(n)} \left(e^{\ln\ln\ln(n) - \ln\ln(n)}\right)^{\frac{3\ln(n)}{\ln\ln(n)}} &= e^{\ln(n)} \left(e^{-3\ln(n) + 3\frac{\ln(n)\ln\ln\ln(n)}{\ln\ln(n)}}\right). \\ &= e^{-2\ln(n) + 3\frac{\ln(n)\ln\ln\ln(n)}{\ln\ln(n)}} \end{aligned}$$

Balls into Bins maximum load

Proof of Lemma 5.1 cont'd.

Grouping the $\ln(n)$'s in the exponents, and evaluating, we have

$$e^{-2\ln(n)} \cdot e^{3\frac{\ln(n)\,\ln\ln\ln(n)}{\ln\ln(n)}} \;=\; n^{-2} \cdot n^{3\frac{\ln\ln\ln(n)}{\ln\ln(n)}}.$$

Balls into Bins maximum load

Proof of Lemma 5.1 cont'd.

Grouping the $\ln(n)$'s in the exponents, and evaluating, we have

$$e^{-2\ln(n)} \cdot e^{3\frac{\ln(n)\ln\ln\ln(n)}{\ln\ln(n)}} = n^{-2} \cdot n^{3\frac{\ln\ln\ln(n)}{\ln\ln(n)}}.$$

If we take n "sufficiently large" ($n \geq e^{e^{e^4}}$ will do it), then $\frac{\ln\ln\ln(n)}{\ln\ln(n)} \leq 1/3$, hence the probability that *some* bin has $\geq M$ balls is at most

$$\frac{1}{n}.$$

□

Balls into Bins maximum load

Proof of Lemma 5.1 cont'd.

Grouping the $\ln(n)$'s in the exponents, and evaluating, we have

$$e^{-2\ln(n)} \cdot e^{3\frac{\ln(n)\ln\ln\ln(n)}{\ln\ln(n)}} = n^{-2} \cdot n^{3\frac{\ln\ln\ln(n)}{\ln\ln(n)}}.$$

If we take n "sufficiently large" ($n \geq e^{e^{e^4}}$ will do it), then $\frac{\ln\ln\ln(n)}{\ln\ln(n)} \leq 1/3$, hence the probability that *some* bin has $\geq M$ balls is at most

$$\frac{1}{n}.$$

□

We can also derive an essentially matching lower pound (using "Poisson approximation"), to show that "with high probability" there will be a bin with $\Omega(\frac{\ln(n)}{\ln\ln(n)})$ balls in it.
We will not prove this (see section 5.3-5.4 of Chapter 5 of [MU]).

Application to Hashing

- An "ideal" hash function should behave like a random function $f : [m] \to [n]$.
- Much research has been done on developing "good" hash functions that "appear" random.
- If we simply assume the hash function behaves randomly, we have precisely the balls-in-bins model.
- Maximum load tells us the maximum number of inputs that hash to the same value. This also defines the limit of the lookup time when we hash a new value.

The power of two choices

Instead of throwing balls randomly, we throw them sequentially with the following tweak: for each ball, we pick two random choices of bins (two different idealized hash functions), and choose the one with the lower load.

The power of two choices

Instead of throwing balls randomly, we throw them sequentially with the following tweak: for each ball, we pick two random choices of bins (two different idealized hash functions), and choose the one with the lower load.

Surprisingly, the maximum load in this case is $\frac{\ln \ln n}{\ln 2} \pm O(1)$ with probability $1 - o(1/n)$.

Note the load reduces from $\Theta\left(\frac{\ln n}{\ln \ln n}\right)$ to $\Theta\left(\ln \ln n\right)$.

The power of two choices

Instead of throwing balls randomly, we throw them sequentially with the following tweak: for each ball, we pick two random choices of bins (two different idealized hash functions), and choose the one with the lower load.

Surprisingly, the maximum load in this case is $\frac{\ln \ln n}{\ln 2} \pm O(1)$ with probability $1 - o(1/n)$.

Note the load reduces from $\Theta\left(\frac{\ln n}{\ln \ln n}\right)$ to $\Theta(\ln \ln n)$.

More generally, if we have $d \geq 2$ choices, the resulting maximum load is $\frac{\ln \ln n}{\ln d} \pm O(1)$ with probability $1 - o(1/n)$.

This is Theorem 17.1 of [MU] (details in Section 17.1/17.2).

Chapter 17 also discusses Cuckoo Hashing, a clever variation of 2-choice hashing, which has been highly successful in practice.

But we do not expect you to know the content of Chapter 17.

References

- Sections 5.1, 5.2 of “Probability and Computing” [MU].