CS 466/666 : Algorithm Design and Analysis. Spring 2018. Waterloo.

Lecture 3: Balls and Bins

Balls and bins is a basic random process that underlies several common phenomenons.

Today we study some standard questions and one interesting variant referred as the power of two choices.

Next time we are going to see some applications in hashing

Balls and Bins

We have m balls and n bins. Each ball is thrown to a uniformly random bin independently.

We would like to study what does a typical situation look like.

There are many questions one can ask to understand the distribution of the balls into the bins.

We start from some easy calculations.

Expected Number of Balls in a Bin

Let $B_{i,j}$ be the indicator variable that ball j is in bin i.

Then $E[\text{\# balls in bin } i] = E\left[\sum_{j=1}^{m} B_{i,j}\right] = \sum_{j=1}^{m} E[B_{i,j}] = \sum_{j=1}^{m} Pr[\text{ball } j \text{ in bin } i] = \sum_{j=1}^{m} \frac{1}{n} = \frac{m}{n}$.

In particular, when $m=n$, the expected number of balls in a bin is one.

Do we expect that most bins have about one ball most of the time?

What is the probability that there is exactly one ball in each bin?

Expected number of empty bins

Let Y_i be the indicator variable that bin i is empty.

Then, $E[Y_i] = Pr(\text{ bin } i \text{ is empty}) = \left(1-\frac{1}{n}\right)^m \approx e^{-\frac{m}{n}}$ (using $1-x \leq e^{-x}$ and $1-x \approx e^{-x}$ for small x)

So, $E[\text{\# of empty bins}] = E\left[\sum_{i=1}^{n} Y_i\right] = \sum_{i=1}^{n} E[Y_i] = n \cdot e^{-\frac{m}{n}}$.

When $m=n$, we expected to see that a $\frac{1}{e}$-fraction of the bins are empty.

Just looking at the expectation of a more "global" variable gives a better understanding of the distribution.

Maximum Load   Question: What is the maximum number of balls in a bin typically?

A simpler question is for what m do we expect to see two balls in a bin (a "collison" occurs).

The _birthday paradox_ is the case when $n=365$ (ignoring Feb 29).

The probability that there are no collision in the first m balls is :

$$\left(1-\tfrac{1}{n}\right)\left(1-\tfrac{2}{n}\right)\cdots\left(1-\tfrac{m-1}{n}\right) \leq e^{-\tfrac{1}{n}}e^{-\tfrac{2}{n}}\cdots e^{-\tfrac{m-1}{n}} = e^{-\tfrac{(m-1)m}{2n}} \approx e^{-\tfrac{m^2}{2n}}.$$

This probability would be smaller than $1/2$ when $m = \sqrt{2n\ln 2}$.

For $n=365$, it says that when $m \geq 22.49$, the probability that the maximum load is at least two is at least $1/2$. This estimate is very close to the exact answer.

To summarize, we expect to see a collision when $m = \theta(\sqrt{n})$ This observation is useful in different places (e.g. hashing, analyzing a heuristic for factoring integers, etc.)

An intuitive explanation is that there are m^2 pairs of possible collisions, and we expect some collision would occur when $m^2 \approx n$, instead of the incorrect intuition that collisions would occur only when $m \approx n/2$.

The maximum load when $m=n$

The probability that a bin has at least k balls is at most $\binom{n}{k}\left(\tfrac{1}{n}\right)^k$, by a union bound.

It is often that we have to deal with binomial coefficients.

Some useful bounds are : $\left(\tfrac{n}{k}\right)^k < \binom{n}{k} < \tfrac{n^k}{k!} < \left(\tfrac{ne}{k}\right)^k$. The proofs are left as exercises.

Using this bound, the above probability is at most $\left(\tfrac{ne}{k}\right)^k \left(\tfrac{1}{n}\right)^k = \tfrac{e^k}{k^k}$.

By the union bound, $\Pr[\text{ some bin has at least } k \text{ balls }] \leq n \cdot \tfrac{e^k}{k^k} = e^{\ln n + k - k\ln k}$.

We would like to estimate the smallest k such that this probability is small enough.

In other words, we want the minimum k such that $k\ln k > \ln n$.

Setting $k = 3\ln n / \ln\ln n$ would do (simple calculations, see MU 5.2.1).

Therefore, with high probability, the maximum load is at most $O(\ln n / \ln\ln n)$.

This $O(\ln n / \ln\ln n)$ comes up in hashing and also in analysis of approximation algorithms
   (e.g. it is the best known approximation ratio of minimizing congestion in a network.)

Coupon Collector   Question: For what m do we expect to have no empty bins?

Let X be the number of balls thrown until there are no empty bins.

Let X_i be the number of balls thrown when there are exactly i empty bins.

Then $E[x] = E[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} E[X_i]$.

How to compute $E[X_i]$? Note that each X_i is a geometric random variable with parameter $p = \frac{i}{n}$.

Recall that a geometric random variable Y is given by the distribution that $Pr(Y=k) = (1-p)^{k-1} p$.

In words, Y is the number of trials until the first success, when the success probability is p.

The expected value of a geometric random variable Y with parameter p is $\frac{1}{p}$.

There are at least three ways to see it:

① direct calculation from the definition with a differentiation trick.

② use $E[Y] = \sum_{i=1}^{\infty} i \, Pr(Y=i) = \sum_{i=1}^{\infty} Pr(Y \geq i) = \sum_{i=1}^{\infty} (1-p)^{i-1} = 1/(1-(1-p)) = \frac{1}{p}$.

  change of summation

③ use conditional probability to argue that $E[Y] = p + (1-p)(E[Y]+1) = (1-p)E[Y] + 1$, hence $E[Y] = 1/p$.

Anyway, we have $E[X] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} 1/(i/n) = \sum_{i=1}^{n} \frac{n}{i} \approx n \ln n$ $\left(\sum_{i=1}^{n} \frac{1}{i} = H_n \text{, the n-th harmonic number} \right)$.

This $n \ln n$ serves as a lower bound for different things. e.g. the cover time of random walks in a complete graph, number of edges needed in graph sparsification by random sampling type algorithms.

Heuristic Arguments

We showed that the maximum load is $O(\ln n / \ln \ln n)$ with high probability. Is it tight?

What is the probability of having an empty bin after throwing $n \ln n + cn$ balls?

The main technical difficulty in analyzing balls and bins is that the random variables involved are not independent, and so for example Chernoff bounds can not be directly applied.

In some situation, like analyzing the number of empty bins, we observe that the events that two bins are nonempty are negatively correlated, and thus Chernoff bounds still apply.

In the following, we pretend that the variables are independent and come up with some heuristic bounds, and later mention that these arguments can be made precise by a method called "Poisson approximation".

Maximum Load

Let p_r be the probability that a bin has exactly r balls.

Then $p_r = \binom{m}{r} \left(\frac{1}{n}\right)^r \left(1 - \frac{1}{n}\right)^{m-r} = \frac{1}{r!} \cdot \frac{m(m-1)\cdots(m-r+1)}{n^r} \left(1 - \frac{1}{n}\right)^{m-r}$

Assuming $m = n \gg r$. Then $p_r = \frac{1}{r!} \frac{n(n-1)\cdots(n-r+1)}{n^r} \left(1-\frac{1}{n}\right)^{n-r} \approx \frac{1}{r!} \cdot 1 \cdot e^{-1} = \frac{1}{er!}$.

We further assume that all bins are independent (while not true, intuitively not too far off).

Then no bin has exactly r balls is at most $\left(1-\frac{1}{er!}\right)^n \le e^{-n/(er!)}$.

If this probability is very small, e.g. $e^{-n/(er!)} \le n^{-2}$, then with high probability there will be some bin with at least r balls.

For $e^{-n/(er!)} \le n^{-2}$ to hold, it suffices to set $-n/(er!) \le -2\ln n \iff r! \le n/(2e\ln n)$

$$\iff \ln r! \le \ln n - \ln\ln n - \ln(2e). \quad (*)$$

By Stirling's approximation that $r! \le e\sqrt{r}\left(\frac{r}{e}\right)^r \le r\left(\frac{r}{e}\right)^r$, (see MU Lemma 5.8 for the first inequality)

$\ln(r!) \le r\ln r - r + \ln r$ $\quad$ (using $\ln(r!) = \sum_{i=1}^{r} \ln i \approx \int_1^r \ln x \, dx = x(\ln x - 1)\Big|_1^r = r\ln r - r$).

Set $r = \ln n / \ln\ln n$.

Then $\ln r! \le \frac{\ln n}{\ln\ln n}(\ln\ln n - \ln\ln\ln n) - \frac{\ln n}{\ln\ln n} + (\ln\ln n - \ln\ln\ln n)$

$\le \ln n - \ln n/\ln\ln n$ $\quad$ (since the sum of the remaining three terms is less than zero)

$\le \ln n - \ln\ln n - \ln(2e)$.

This shows that $(*)$ holds when $r = \ln n / \ln\ln n$, and therefore there exists some bin with load $\Omega(\ln n / \ln\ln n)$.

Coupon Collector

To estimate the probability that some bin is empty after $n\ln n + cn$ balls, again we use

$$p_r = \frac{1}{r!} \frac{m(m-1)\cdots(m-r+1)}{n^r}\left(1-\frac{1}{n}\right)^{m-r} \approx \frac{1}{r!}\left(\frac{m}{n}\right)^r e^{-m/n} \quad \text{when} \quad m, n \gg r.$$

For $m = n\ln n + cn$, $p_0 \approx e^{-c}/n$.

So, the probability of having some empty bin is $\approx 1 - \left(1-\frac{e^{-c}}{n}\right)^n \approx 1 - e^{-e^{-c}} = 1 - \frac{1}{e^{\frac{1}{e^c}}}$.

When c is a large positive constant, this is very close to zero.

When c is a large negative constant, this is very close to one.

This is a "sharp" threshold phenomenon, for which we expect the event happens when there are very close to $n\ln n$ balls.

Poisson Approximation (optional)

Why can we assume independence in previous arguments?

No, we can not, but we can make it precise by using the Poisson approximation technique.

Recall that $p_r = \binom{m}{r}\left(\frac{1}{n}\right)^r\left(1-\frac{1}{n}\right)^r \approx e^{-m/n}(m/n)^r/r!$ for small r.

Think of m/n as the mean.

Define a __Poisson__ random variable with parameter μ by the probability distribution $Pr(x=j) = e^{-\mu}\mu^j/j!$.

Then p_r is just a Poisson random variable with $\mu = m/n$.

Note that it is a probability distribution, with expected value μ, and it is a good approximation of

  binomial random variables (MU Thm 5.5).

Let $X_i^{(m)}$ be the number of balls in bin i when m balls are thrown, and $Y_i^{(m)}$ be a Poisson random

  variable with mean m/n.

A main difference between the distributions $(X_1^{(m)}, X_2^{(m)}, ..., X_n^{(m)})$ and $(Y_1^{(m)}, Y_2^{(m)}, ..., Y_n^{(m)})$ is that

  $\sum_i Y_i^{(m)}$ may not be equal to m.

There are two key points in the proof:

① Conditioned on $\sum_i Y_i^{(m)} = m$. Then the two distributions are the same. ([MU, Theorem 5.6])

② $\sum_i Y_i^{(m)} = m$ happens with reasonable probability. i.e. with probability at least $\frac{1}{e\sqrt{m}}$. ([MU, Theorem 5.7])

Combining these two points, if we can give a good upper bound in the Poisson distribution, we can give a just

  slightly bigger upper bound in the original distribution (i.e. just a factor $e\sqrt{m}$ bigger).

For the maximum load problem and the coupon collector problem, we can give a very small upper bound on

  the bad events in the Poisson distribution, the heuristic arguments can be made precise.

The key point is that we can work with __independent__ Poisson random variables, and apply

  the corresponding Chernoff bound to show that bad events happen with very small probability.

Power of Two Choices (optional)

Now we know that when n balls are thrown into n bins. Then the maximum load is $\Theta(\ln/\ln\ln n)$ w.h.p.

Consider the following variant when each ball is thrown we pick two random bins and put the ball in the

  bin with fewer balls.

Surprisingly this simple modification can significantly reduce the maximum load to $O(\ln\ln n)$!

The intuition is simple. A ball is of height i if it is the i-th ball put in the bin. Suppose we can

  bound the number of bins with at least i balls by β_i, over the entire course of the process. What

should be β_{i+1}? A ball is of height $i+1$ if the two random bins both have at least i balls. This happens with probability at most $(\beta_i/n)^2$. Hence $\frac{\beta_{i+1}}{n} \leq (\frac{\beta_i}{n})^2$. Solving the recurrence gives that β_j becomes $O(\ln n)$ when $j = O(\ln \ln n)$. At this point the number is too small to apply concentration inequalities for the induction, but it is easy to finish the proof from there.

We will use the following Chernoff bound

$$Pr(B(n,p) \geq 2np) \leq e^{-np/3}, \text{ where } B(n,p) \text{ is the binomial random variable with } n \text{ trials and success prob. } p.$$

Let $\beta_4 = n/4$ and $\beta_{i+1} = 2\beta_i^2/n$.

Let ε_i be the event that after all n balls are thrown the number of bins with at least i balls is $\leq \beta_i$.

Note that β_4 holds with probability 1.

We will prove that if ε_i holds then ε_{i+1} holds with high probability (until β_i becomes too small).

In the following we condition on the event ε_i.

Let $Y_t = 1$ if the t-th ball has height at least $i+1$.

Then $Pr(Y_t = 1) \leq \beta_i^2/n$.

Let $p_i = \beta_i^2/n$. Then $Pr(\sum_{t=1}^{n} Y_t > k) \leq Pr(B(n,p_i) > k)$.

So $Pr((\# \text{ bins with at least } i+1 \text{ balls}) > k) \leq Pr((\# \text{ balls with height at least } i+1) > k \mid \varepsilon_i)$

$$= Pr(\sum_{t=1}^{n} Y_t > k \mid \varepsilon_i)$$

$$\leq Pr(B(n,p_i) > k \mid \varepsilon_i)$$

Set $k = \beta_{i+1} = 2np_i$, then the above probability $\leq \frac{Pr(B(n,p_i) > 2np_i)}{Pr(\varepsilon_i)} \leq \frac{1}{Pr(\varepsilon_i) \cdot e^{p_i \cdot n/3}}$ by Chernoff.

This implies that $Pr(\neg\varepsilon_{i+1} \mid \varepsilon_i) \leq \frac{1}{n^2 Pr(\varepsilon_i)}$ as long as $p_i \cdot n \geq 6\ln n$.

So $Pr(\neg\varepsilon_{i+1}) = Pr(\neg\varepsilon_{i+1} \mid \varepsilon_i) Pr(\varepsilon_i) + Pr(\neg\varepsilon_{i+1} \mid \neg\varepsilon_i) Pr(\neg\varepsilon_i)$

$$\leq Pr(\neg\varepsilon_{i+1} \mid \varepsilon_i) Pr(\varepsilon_i) + Pr(\neg\varepsilon_i) \leq \frac{1}{n^2} + Pr(\neg\varepsilon_i) \text{ as long as } p_i \cdot n \geq 6\ln n.$$

To finish the proof we need two more steps. First is to prove that $p_i \cdot n < 6\ln n$ in $O(\ln \ln n)$ steps. And second is to handle the case when $p_i \cdot n < 6\ln n$.

The first step is easy. A simple induction can show that $\beta_{i+4} \le n/2^{2^i}$ and therefore $p_i \cdot n < 6\ln n$ in $O(\ln\ln n)$ steps. And thus $\Pr(\neg \mathcal{E}_i) \le O(\ln\ln n)/n^2$ in this step.

The second step is also easy. By Chernoff bound we can show that whp there are at most $O(\ln n)$ bins with at least $\Omega(\ln\ln n)$ balls at this stage. Then those bins are just too few that we can finish the argument by naive bound ($(\frac{\ln n}{n})^2$ to reach one step higher and there are at most $\binom{n}{2}$ ways of choosing two balls) to show that there are at most 2 bins with one more ball. And then no more bin with one more ball.

This concludes the proof (sketch).

<u>Remark:</u> The bound is tight. One can show that whp there is a bin with $\Omega(\ln\ln n)$ balls. What if we choose d random bins in each step and put the ball in a least loaded bin?

References

Balls and bins is from chapter 5 of "probability and computing" by Mitzenmacher and Upfal. You can read the details about Poisson approximation there.

Power of two choices is from chapter 14 of the same book. It is a well-studied topic with applications to hashing and further generalizations (e.g. balanced allocations on graphs).

The book by Dubhashi and Panconesi on "Concentration of measure for the analysis of randomized algorithms" is an excellent resource about concentration inequalities and their algorithmic applications. For example, you can read about negative correlations and also the theory of martingales there.