# CS 466/666 Algorithm Design and Analysis . Spring 2022

## Lecture 4: Balls and Bins

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

Today we study some basic 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 $E[\#\text{ balls in bin } i] = E[\sum_{j=1}^{m} B_{i,j}] = \sum_{j=1}^{m} E[B_{i,j}] = \sum_{j=1}^{m} \Pr[\text{ball } j \text{ in bin } i] = \sum_{j=1}^{m} \frac{1}{n} = \frac{m}{n}$.

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

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

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.

In fact, this random variable can be shown to be concentrated around the mean. so we do expect to see around $1/e$-fraction of empty bins when $m=n$.

---

**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 <u>birthday paradox</u> is the case when $n=365$ (ignoring Feb 29).

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

$$(1-\frac{1}{n})(1-\frac{2}{n})\dots(1-\frac{m-1}{n}) \le e^{-\frac{1}{n}} e^{-\frac{2}{n}} \dots e^{-\frac{m-1}{n}} = e^{-\frac{(m-1)m}{2n}} \approx e^{-\frac{m^2}{2n}}.$$

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

For $n=365$, it says that when $m \ge 22.49$, 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 collison 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 collison would occur when $m^2 \approx n$, instead of the incorrect intuition that collisons 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 $\binom{n}{k}(\frac{1}{n})^k$, by a union bound.

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

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

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

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

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 still the best known approximation ratio for congestion minimization).

---

**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 differentition 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 $E[X] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} 1/(i/n) = \sum_{i=1}^{n} \frac{n}{i} \approx n \ln n$ ($\sum_{i=1}^{n} \frac{1}{i} = H_n$, the n-th harmonic number).

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 type bounds can not be directly applied.

In some situation, such as analyzing the number of empty bins, we observe that the events that two bins are nonempty are negatively correlated, and thus Chernoff type 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 $p_r = \binom{m}{r}\left(\frac{1}{n}\right)^r \left(1-\frac{1}{n}\right)^{m-r} = \frac{1}{r!} \frac{m(m-1)\cdots(m-r+1)}{n^r} \left(1-\frac{1}{n}\right)^{m-r}$

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 $\left(1-\frac{1}{er!}\right)^n \leq e^{-n/(er!)}$.

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

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

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

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)

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

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

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

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

## Coupon Collector

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

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

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

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

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 $\mu$ 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 $\mu$, 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 $(X_1^{(m)}, X_2^{(m)}, \ldots, X_n^{(m)})$ and $(Y_1^{(m)}, Y_2^{(m)}, \ldots, Y_n^{(m)})$ 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).

More precisely, suppose we prove that an event $\mathcal{E}$ happens in the Poisson setting with probability $p$. then we can use the above claims to conclude that

$$p \geq Pr_Y(\mathcal{E}) \geq Pr_Y(\mathcal{E} \mid \textstyle\sum_i Y_i = m) Pr(\sum_i Y_i = m) \overset{②}{\geq} \frac{1}{e\sqrt{m}} Pr_Y(\mathcal{E} \mid \textstyle\sum_i Y_i = m) \overset{①}{=} \frac{1}{e\sqrt{m}} Pr_X(\mathcal{E}) \Rightarrow Pr_X(\mathcal{E}) \leq e\sqrt{m}\,p.$$

For the maximum load and the coupon collector problem, we can prove a very small upper bound on the bad event in the independent Poisson setting (i.e. $Pr_Y(\mathcal{E}) \leq p$ for very small $p$), and so this translates into a small upper bound on the bad event probability in the balls and bins setting as well.

The key point of this technique is that we can work with <u>independent</u> random variables, and e.g. we can apply Chernoff bounds to show that bad events happen with small probability.

The proofs are very nice mathematically, but the main reason that this is optional is because the technique seems to be tailored made to the analysis of balls and bins but does not apply in other settings.

---

## <u>Power of Two Choices</u> (optional)

Now we know that when $n$ balls are thrown into $n$ bins. Then the maximum load is $\Theta(\ln n / \ln\ln n)$ 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 $\beta_i$, over the entire course of the process. What should be $\beta_{i+1}$? 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 $\frac{\beta_{i+1}}{n} \leq \left(\frac{\beta_i}{n}\right)^2$. Solving the recurrence gives that $\beta_j$ 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:

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

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

Let $\mathcal{E}_i$ 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 $\beta_4$ holds with probability 1.

We will prove that if $\mathcal{E}_i$ holds then $\mathcal{E}_{i+1}$ holds with high probability (until $\beta_i$ becomes too small).

In the following we condition on the event $\mathcal{E}_i$.

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 $\Pr\left(\sum_{t=1}^{n} Y_t > k\right) \leq \Pr(B(n, p_i) > k)$.

So $\Pr\left((\#\text{ bins with at least } i+1 \text{ balls}) > k\right) \leq \Pr\left((\#\text{ balls with height at least } i+1) > k \mid \mathcal{E}_i\right)$

$$= \Pr\left(\sum_{t=1}^{n} Y_t > k \mid \mathcal{E}_i\right)$$

$$\leq \Pr(B(n, p_i) > k \mid \mathcal{E}_i)$$

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

This implies that $\Pr(\neg\mathcal{E}_{i+1} \mid \mathcal{E}_i) \leq \dfrac{1}{n^2 \Pr(\mathcal{E}_i)}$ as long as $p_i n \geq 6\ln n$.

So $\Pr(\neg\mathcal{E}_{i+1}) = \Pr(\neg\mathcal{E}_{i+1} \mid \mathcal{E}_i)\Pr(\mathcal{E}_i) + \Pr(\neg\mathcal{E}_{i+1} \mid \neg\mathcal{E}_i)\Pr(\neg\mathcal{E}_i)$

$\leq \Pr(\neg\mathcal{E}_{i+1} \mid \mathcal{E}_i)\Pr(\mathcal{E}_i) + \Pr(\neg\mathcal{E}_i) \leq \frac{1}{n^2} + \Pr(\neg\mathcal{E}_i)$ as long as $p_i n \geq 6\ln n$.

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

The first step is easy. A simple induction can show that $\beta_{i+4} \leq n/2^{2^i}$ and therefore $p_i n < 6\ln n$ in $O(\ln\ln n)$ steps. And thus $\Pr(\neg\mathcal{E}_i) \leq O(\ln\ln n)/n^2$ 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 with $\Omega(\ln\ln n)$ balls at this stage. Then those bins are just too few that we can finish the argument by simple bound ($(\ln n/n)^2$ 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).

Remark: 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 are from chapter 5 of [MU]. You can read the details of Poisson approximation there.

Power of two choices is from chapter 14 of [MU]. It is a well-studied topic with applications in 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. You can read about negative correlations and the theory of martingales there.