Balls and Bins

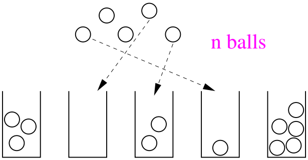
Josep Díaz Maria J. Serna Conrado Martínez
U. Politècnica de Catalunya

RA-MIRI 2023–2024

Balls and Bins

Basic Model: Given n balls, we throw each one independently and uniformly into a set of m bins.

$$\mathbb{P}[\text{ball } i \to \text{bin } j] = \frac{1}{m}.$$

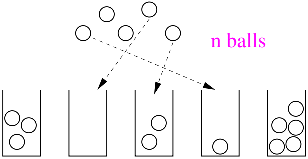


Probability space: $\Omega = \{(b_1, b_2, \ldots, b_n)\}$ where $b_i \in \{1, \ldots, m\}$ denotes the index of the bin containing the i-th ball: $|\Omega| = m^n$. For any $w \in \Omega$, $\mathbb{P}[w] = (\frac{1}{m})^n$

Balls and Bins

Basic Model: Given n balls, we throw each one independently and uniformly into a set of m bins.

$$\mathbb{P}[\text{ball } i \to \text{bin } j] = \frac{1}{m}.$$



Probability space: $\Omega = \{(b_1, b_2, \ldots, b_n)\}$ where $b_i \in \{1, \ldots, m\}$ denotes the index of the bin containing the i-th ball: $|\Omega| = m^n$. For any $w \in \Omega$, $\mathbb{P}[w] = (\frac{1}{m})^n$

Balls and Bins as a model

Balls and Bins models are very useful in different areas of computer science. For ex.:

- The hashing data structure: the keys are the balls and the slots in the array are the bins.
- Many situations in routing in nets: the balls represent the connectivity requirements and the bins the paths in the network
- Load balancing randomized algorithms, the balls are the jobs and the bins are the servers.

Example

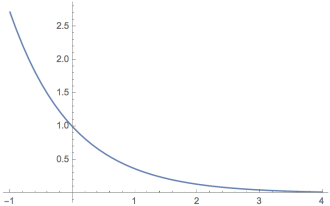
Recall that, as an application of Chernoff bounds, we proved that for n balls (jobs) and m bins (servers), under a uniform and independent distribution of jobs to servers, for $n \gg m$, the probability the load of a server deviates from the expected load n/m is $\leqslant 1/m^3$.

General rules for the analysis of Balls & Bins

n balls to m bins.

- X_j is the random variable counting the number of balls into bin j. Then $X_j \sim \text{Bin}(n, \frac{1}{m})$.
- As we know: $X_1, \ldots X_m$ are not independent.
- The average load in a bin is $\mu = \mathbb{E}[X_j] = n/m$.
- Rule of thumb to do the analysis:
  - If $n \gg m$, (μ large) use Chernoff bounds,
  - if $n = \Theta(m)$, ($\mu \in \Theta(1)$), use the Poisson approximation.

Recall that for very small x,
$e^x \sim 1 + x$
$e^{-x} \sim 1 - x$.

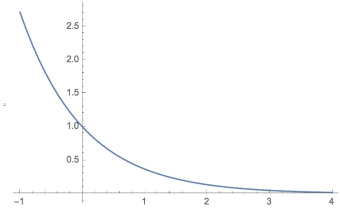


General rules for the analysis of Balls & Bins

n balls to m bins.

- X_j is the random variable counting the number of balls into bin j. Then $X_j \sim \text{Bin}(n, \frac{1}{m})$.
- As we know: $X_1, \ldots X_m$ are not independent.
- The average load in a bin is $\mu = \mathbb{E}[X_j] = n/m$.
- Rule of thumb to do the analysis:
  - If $n \gg m$, (μ large) use Chernoff bounds,
  - if $n = \Theta(m)$, ($\mu \in \Theta(1)$), use the Poisson approximation.

Recall that for very small x,
$e^x \sim 1 + x$
$e^{-x} \sim 1 - x$.



The Poisson Distribution

Recall that for $X \sim \text{Bin}(n, p)$, for large n and small p, we can have a good approximation: $\mathbb{P}[X = k] = \frac{e^{-\lambda}\lambda^k}{k!}$, where $\lambda = \mathbb{E}[X] = \mu = pn$.

The Poisson Distribution: Basic Properties

Assume that $Y \sim \text{Poisson}(\lambda)$ approximates $X \sim \text{Bin}(n, p)$, then as $\mathbb{E}[X] = np$ seems natural that $\mathbb{E}[Y] = np = \lambda$. On the other hand $\mathbb{V}[X] = np(1 - p) = \lambda(1 - p)$ and if p is small $\mathbb{V}[X] \sim \lambda$ and $\mathbb{V}[Y] = \lambda$.

Sum of Poisson r. v.

Lemma

If $Y \sim$ *Poisson*(λ) *and* $Z \sim$ *Poisson*(λ') *are independent, then* $Y + Z \sim$ *Poisson*$(\lambda + \lambda')$*.*

Proof

$$\mathbb{P}[Y + Z = j] = \sum_{k=0}^{j} \mathbb{P}[(Y = k) \cap (Z = j - k)] = \sum_{k=0}^{j} \frac{e^{-\lambda} e^{-\lambda'} \lambda^k \lambda'^{j-k}}{k!(j-k)!}$$

$$= \frac{e^{-(\lambda+\lambda')}}{j!} \sum_{k=0}^{j} \frac{j!}{k!(j-k)!} \lambda^k \lambda'^{j-k} = \frac{e^{-(\lambda+\lambda')}}{j!} \sum_{k=0}^{j} \binom{j}{k} \lambda^k (\lambda')^{j-k}$$

$$= \frac{e^{-(\lambda+\lambda')} \times (\lambda + \lambda')^j}{j!} \Rightarrow (Y + Z) \sim \text{Poisson}(\lambda + \lambda')$$

$\square$

Basic facts

Recall X_j counts the number of balls in the j-th bin.

- Probability all n balls go to the same bin: $(\frac{1}{m})^n$.
- Probability that bin j is empty:
  $\mathbb{P}\left[X_j = 0\right] = (1 - \frac{1}{m})^n \sim e^{-\frac{n}{m}} = e^{-\lambda}$.
- Let Y be the number of empty bins, $\mathbb{E}[Y]$?.
  For $1 \leqslant j \leqslant m$, let Y_j be the r.v.defined as $Y_j = 1$ iff bin j is empty, 0 otherwise. Then,
  $\mathbb{E}[Y] = \sum_{j=1}^{m} \mathbb{E}\left[Y_j\right] = \sum_{j=1}^{m} \mathbb{P}\left[X_j = 0\right] = m(1 - 1/m)^n$. So, the expected number of empty bins is

$$\mathbb{E}[Y] \sim me^{-\lambda}.$$

Probability the j-th bin contains 1 ball

We can assume that m and n are large, (so $p = 1/m$ is small), $\lambda = n/m = \Theta(1)$

Exact computation: $\mathbb{P}\left[X_j = 1\right] = \binom{n}{1}(1/m)^1(1-1/m)^{n-1}$, where $\binom{n}{1}$ is the number of choices were exactly 1 ball goes into bin j,

$(1-1/m)^{n-1}$: remaining balls do not go to bin j.

$\mathbb{P}\left[X_j = 1\right] = \frac{n}{m}(1-1/m)^n(1-1/m)^{-1}$

Poisson approximation: Taking $\lambda = \frac{n}{m}$ and $(1-1/m)^n \sim e^{-\lambda}$ and noticing $(1-1/m) \to 1$:

$$\mathbb{P}\left[X_j = 1\right] \sim \lambda e^{-\lambda}.$$

Example

For $n = 3000$ and $m = 1000$, $\lambda = 3$, the exact value of $\mathbb{P}[X_i = 1] = 0.149286$ and the Poisson approximation is 0.149361.

Probability the j-th bin contains exactly r balls

Assume that m and n are large and $n, m > r$

Exact computation: $\mathbb{P}[X_j = r] = \binom{n}{r}(1/m)^r(1-1/m)^{n-r}$.

Poisson approximation:

$(1-1/m)^{n-r} = (1-1/m)^n(1-1/m)^{-r} = e^{-\lambda} \cdot 1^{-r}$

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

$$\mathbb{P}[X_j = r] \sim \frac{\lambda^r e^{-\lambda}}{r!}$$

Example

For $n = 4000$ and $m = 2000$, $\lambda = 2$, and $r = 100$, the exact value of $\mathbb{P}[X_i = r] = 5.54572 \times 10^{-130}$ and the approximation is 1.83826×10^{-130}

Probability of collisions

$\mathbb{P}[\text{at least 1 bin has more than 1 ball }] =$
$1 - \mathbb{P}\left[\text{every bin } j \text{ has } X_j \leqslant 1\right]$. If $k-1$ balls went to $k-1$ different bins. Then,

$$\mathbb{P}[\text{The } k\text{th. ball goes into a non-empty bin}] = \frac{k-1}{m}$$

$$\mathbb{P}[\text{The } k\text{th. ball goes into an empty bin}] = (1 - \frac{k-1}{m})$$

$$\mathbb{P}[\text{every ball goes to an empty bin}] = \prod_{i=1}^{n}\left(1 - \frac{i-1}{m}\right)$$

$$= \prod_{i=1}^{n-1}\left(1 - \frac{i}{m}\right) \sim \prod_{i=1}^{n} e^{-i/m}$$

$$= e^{-\sum_{i=1}^{n-1} i/m} = e^{-\frac{n(n-1)}{2m}} \sim e^{-\frac{n^2}{2m}}$$

Therefore, $\mathbb{P}[\text{at least 1 bin } i \text{ has } X_i > 1] \sim 1 - e^{-\frac{n^2}{2m}}$.

Birthday problem

Example

How many students should be in a class in order to have that, with probability $>$ $1/2$, at least 2 have the same birthday?

This is the same problem as above, with $m = 365$:

We need $e^{-\frac{n^2}{2m}} \leqslant \frac{1}{2} \Rightarrow \frac{n^2}{2m} \leqslant \ln 2 \sim 0.69$
$\Rightarrow n = \sqrt{2m \ln 2}$. If $m = 365$ then $n = 22.49$.

If there are more than 23 students in a class, with probability greater than $1/2$, two or more students will have the same birthday.

Coupon Collector's problem

How many balls do we need to throw to assure that w.h.p. every bin contains $\geqslant 1$ balls?

- Let Y a r.v. counting the number of balls we have to throw until having no empty bins
- For $1 \leqslant i \leqslant m$, let $Y_i =$ # balls thrown since the moment in which $i-1$ bins are not empty until a ball goes into an empty bin.
- $Y_1 = 1$ and $Y = \sum_{i=1}^{m} Y_i$.
- $\mathbb{P}[\text{new ball into non-empty bin} \,|\, i-1 \text{ non-empty bins}] = \frac{i-1}{m}$.
- $\mathbb{P}[\text{new ball into empty bin} \,|\, i-1 \text{ non-empty bins}] = 1 - \frac{i-1}{m}$.

Coupon Collector's problem: $\mathbb{E}[Y]$

$Y_i =$ # of balls we have to throw to hit an empty bin having $i-1$ non-empty

$$\mathbb{P}[Y_i = k] = \left(\frac{i-1}{m}\right)^{k-1} \left(\underbrace{1 - \frac{i-1}{m}}_{p_i}\right).$$

Therefore $Y_i \in \text{Geom}(p_i)$ and $\mathbb{E}[Y_i] = \frac{m}{m+i+1}$.

$$\mathbb{E}[Y] = \sum_{i=1}^{m} \mathbb{E}[Y_i] = \sum_{i=1}^{m} \frac{m}{m-i+1} = m \sum_{j=1}^{m} \frac{1}{j} = m(\ln m + \mathcal{O}(1)).$$

Coupon Collector's problem: Concentration

Let $\mathbb{E}[Y] = \mathcal{O}(m \ln m) \sim cm \ln m$ for constant $c > 1$

- For any bin j, define the event $A_{j,r}$: bin j is empty after the first r throws.
- Notice events $A_{1,r}, A_{2,r}, \ldots A_{m,r}$ are not independent.
- $\mathbb{P}\left[A_{j,r}\right] = (1 - \frac{1}{m})^r \sim e^{-r/m}$
- For $r = cm \ln m \Rightarrow \mathbb{P}\left[A_{j,cm \ln m}\right] \leqslant e^{-cm \ln m/m} = m^{-c}$.
- Let W be a r.v. counting the number of balls needed to make every bin have load $\geqslant 1$.

$$\mathbb{P}[W > cm \lg m] = \mathbb{P}\left[\cup_{j=1}^{m} A_{j,cm \ln m}\right] \underbrace{\leqslant}_{UB} \sum_{j=1}^{m} \mathbb{P}\left[A_{j,cm \ln m}\right]$$

$$\leqslant \sum_{j=1}^{m} m^{-c} = m^{1-c}.$$

Coupon Collector's problem: Concentration Bounds

- The previous bound using UB is more tight than the one using Chebyshev or Chernoff on random variable Y.
- In Section 5.4.1 of MU book, there is a sharper bound for the Coupon collector's, using the Poisson approximation.

Maximum Load

This is a particular case of the job and servers with sharper bounds

Theorem

If we throw n balls independently and uniformly into $m = n$ bins, then the maximum load of a bin is at most $\left(\frac{3 \ln n}{\ln \ln n}\right)$, with probability $\leqslant 1 - \frac{1}{n}$ if n is large enough.

Recall that, if for any bin $1 \leqslant j \leqslant n$, $X_j =$ is a r.v. with its load. We know $\{X_j\}$ are not independent and $\mathbb{E}\left[X_j\right] = n/n = 1$.

To show the above bound we use the following inequality:

$$\binom{n}{k}\frac{1}{n^k} \leqslant \frac{1}{k!} \leqslant \left(\frac{e}{k}\right)^k$$

Max-load: Proof Upper Bound

There are $\binom{n}{k}$ ways to choose k balls out of n and the probability that all them land in bin j is $(1/m)^k = (1/n)^k$, hence for $1 \leqslant k \leqslant n$, $\mathbb{P}\left[X_j \geqslant k\right] \leqslant \binom{n}{k}\frac{1}{n^k} \leqslant (\frac{e}{k})^k$.

We want to prove that for $k \geqslant \frac{3\ln n}{\ln\ln n}$ and n large enough

$$\mathbb{P}\left[\exists j : X_j \geqslant \frac{3\ln n}{\ln\ln n}\right] \leqslant \frac{1}{n}.$$

By the union bound and since $k \geqslant 3\ln n/\ln\ln n$

$$\begin{aligned}
\mathbb{P}\left[\exists j : X_j \geqslant k\right] \leqslant n\left(\frac{e}{k}\right)^k &\leqslant n\left(\frac{e\ln\ln n}{3\ln n}\right)^{3\ln n/\ln\ln n} \\
&< e^{\ln n}\left(\frac{\ln\ln n}{\ln n}\right)^{3\ln n/\ln\ln n} \\
&= e^{\ln n}\left(e^{\ln\ln\ln n - \ln\ln n}\right)^{3\ln n/\ln\ln n} \\
&= e^{\ln n}e^{3\ln n(\ln\ln\ln n/\ln\ln n) - 3\ln n} = e^{-2\ln n}e^{3\ln n\frac{\ln\ln\ln n}{\ln\ln n}} \\
&= n^{-2}e^{3\ln n\frac{\ln\ln\ln n}{\ln\ln n}} \leqslant n^{-2}\cdot o(n) \leqslant n^{-1}, \qquad \text{for large } n.
\end{aligned}$$

Further considerations on Max-load

1. The same proof could be extended to the case of n balls and m bins, with the constrain $n < m \ln m$.
2. We can obtain the same result by using Chernoff's bounds. (Nice exercise!)
3. In fact, the result could be extended to prove the Lower Bound: that w.h.p. the max-load is $\Omega(\frac{\ln n}{\ln \ln(n)})$ balls. One easy way to prove the lower bound is using Chebyshev's bound.
4. That result yields: Throwing n balls to n bins, w.h.p. we have a max-load of $\Theta(\frac{\ln n}{\ln \ln(n)})$.
5. We can obtain sharper bounds for max-load, using strong inequalities (Azuma-Hoeffding) or the Poisson approximation.

Poisson approximation

1. A difficulty with the exact (binomial) Balls & Bin model is that random variables could be dependent (for ex. bin's load).
2. We have seen how to approximate the expressions arising from the exact computations by a Poisson, if p is small and n is large.
3. However, under the right conditions, we can approach the whole solution to the problem by using Poisson r.v. instead of Binomial. In the binomial case we have exactly n balls with probability $p = 1/m$, in the Poisson case we have an intensity $\lambda = n/m$, where n is the expected number of balls being used.
4. The Poisson case is to use independent Poisson random variables. It can be shown, under certain conditions, that the approach gives a good approximation to the solution. See for ex. section 5.4 in MU.