Balls and Bins

RA-MIRI QT Curs 2020-2021
Balls and Bins

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

\[
\Pr [\text{ball } i \rightarrow \text{bin } j] = \frac{1}{m}.
\]

Probability space: \( \Omega = \{(b_1, b_2, \ldots, b_n)\} \) where \( b_i \in \{1, \ldots, m\} \) denotes the index of the bin containing ball \( i \)-th. ball: \(|\Omega| = m^n\).

For any \( w \in \Omega, \Pr [w] = \left(\frac{1}{m}\right)^n\)
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.

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, was \( 1/m^3 \).
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 \( X_j \in B(n, \frac{1}{m}) \).
- As we know: \( X_1, \ldots X_m \) are not independent.
- The average load in a bin is \( \mu = E[X_j] = n/m \).
- Rule of thumb to do the analysis:
  - If \( n \gg m \), (\( \mu \) large) use Chernoff bounds,
  - if \( n = m \), (\( \mu \in \Theta(1) \)), 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 \( X \in B(n, p) \), for large \( n \) and small \( p \), we can have a good approximation: 
\[
\Pr[X = k] = \frac{e^{-\lambda} \lambda^k}{k!},
\]
where \( \lambda = \mathbb{E}[X] = \mu = pn \).

For any \( \lambda \in \mathbb{R}^+ \), a r.v. \( X \) is said to have a Poisson \( P(\lambda) \) distribution, if its PMF is 
\[
p_X(k) = \frac{e^{-\lambda} \lambda^k}{k!}, \text{ for any } k = 0, 1, 2, 3, \ldots
\]

Poisson is one of the most ”natural” distributions: number of typos, number of rain drops in a square meter of roof, etc..
Assume that \( Y \in P(\lambda) \) approximates \( X \in B(n, p) \), then as 
\[ E[X] = np \] seems natural that \( E[Y] = np = \lambda \) and as 
\[ \text{Var}[X] = np(1 - p) = \lambda(1 - p) \] and as \( p \) is small \( \text{Var}[X] \sim \lambda \) and 
\[ \text{Var}[Y] = \lambda. \] Formally, If \( Y \in P(\lambda) \):

- \( E[Y] = \lambda. \)

\[
E[Y] = \sum_{k=0}^{\infty} k \frac{e^{-\lambda} \lambda^k}{k!} = e^{-\lambda} (\lambda + \frac{2\lambda^2}{2!} + \frac{3\lambda^2}{3!} \cdots )
\]

\[
= e^{-\lambda} \lambda (1 + \lambda + \frac{2\lambda^2}{2!} + \frac{3\lambda^2}{3!} \cdots ) = e^{-\lambda} \lambda e^\lambda
\]
Variance of Poisson r.v.

- \( \text{Var} [Y] = \lambda. \)

To prove it, instead of computing \( E [X^2] \) we compute \( E [X(X - 1)]. \)
Notice \( \text{Var} [X] = E [X^2] - E [X]^2 = E [X(X - 1)] + E [X] - E [X]^2. \)

\[
E [X(X - 1)] = \sum_{x=0}^{\infty} x(x - 1) \frac{\lambda^x e^{-\lambda}}{x!} = \sum_{x=2}^{\infty} \frac{\lambda^2 \lambda^{x-2} e^{-\lambda}}{(x-2)!}
\]

\[
= e^{-\lambda} \lambda^2 \sum_{x=2}^{\infty} \frac{\lambda^{x-2}}{(x-2)!} = e^{-\lambda} \lambda^2 \sum_{y=0}^{\infty} \frac{\lambda^y}{(y)!}
\]

\[
= e^{-\lambda} \lambda^2 e^\lambda
\]

So, \( \text{Var} [X] = \lambda^2 + \lambda - \lambda^2 \)
Lemma If $Y \in P(\lambda)$ and $Z \in P(\lambda')$ are independent, then $Y + Z \in P(\lambda + \lambda')$.

Proof

$$
\Pr[Y + Z = j] = \sum_{k=0}^{j} \Pr[(Y = k) \cap (Z = j - k)] = \sum_{k=0}^{j} \frac{e^{-\lambda} e^{-\lambda'} \lambda^k \lambda'^{j-k}}{k!(j-k)!} = \frac{e^{-(\lambda + \lambda')}}{j!} \sum_{k=0}^{j} \frac{j!}{k!(j-k)!} \lambda^k \lambda'^{j-k} = \frac{e^{-(\lambda + \lambda')}}{j!} \sum_{k=0}^{j} \binom{j}{k} \lambda^k (\lambda')^{j-k} = \frac{e^{-(\lambda + \lambda')}}{j!} \times (\lambda + \lambda')^j \Rightarrow (Y + Z) \in P(\lambda + \lambda') \quad \square
$$
Basic facts

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

- Probability all $n$ balls fell in the same bin: $(\frac{1}{m})^n$.
- Probability that bin $j$ is empty:
  \[ \Pr[X_j = 0] = (1 - \frac{1}{m})^n \sim e^{-\frac{n}{m}} = e^{-\lambda}. \]
- Let $Y$ be the number of empty bins, $E[Y]$.
  For $1 \leq j \leq m$, let $Y_j$ be and the r.v. defined as $Y_j = 1$ iff bin $j$ is empty, 0 otherwise. Then,
  \[ E[Y] = \sum_{j=1}^{m} E[Y_j] = \sum_{j=1}^{m} \Pr[X_j = 0] = m(1 - 1/m)^n. \]
  So, the expected number of empty bins is
  \[ E[Y] \sim me^{-\lambda}. \]
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: \( \Pr[X_j = 1] = \binom{n}{1}(1/m)^1(1 - 1/m)^{n-1} \)
where \( \binom{n}{1} \) number choices exactly 1 ball goes into bin $j$,

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

\( \Pr[X_j = 1] = \frac{n}{m}(1 - 1/m)^n(1 - 1/m)^{-1} \)

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

\( \Pr[X_j = 1] \sim \lambda e^{-\lambda} \).

For $n = 3000$ and $m = 1000$, \( \lambda = 3 \), the exact value of \( \Pr[X_i = 1] = 0.149286 \) and the Poisson approximation is 0.149361.
Probability the $j$-th bin contains exactly $r$ balls

We can assume that $m$ and $n$ are large, $n, m > r$.

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

Poisson approximation:

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

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

$$= \frac{1}{r!} \lambda(1 - \frac{1}{n}) \cdots \lambda(1 - \frac{r+1}{n}) = \lambda^r$$

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

For $n = 4000$ and $m = 2000$, $\lambda = 2$, and $r = 100$, the exact value of $\Pr[X_i = r] = 5.54572 \times 10^{-130}$ and the approximation is $1.83826 \times 10^{-130}$.
Probability that at least one bin has a collision

\[ \Pr[ \text{at least 1 bin has more than 1 ball } ] = 1 - \Pr[ \text{every bin } j \text{ has } X_j \leq 1]. \]

If \( k - 1 \) balls went to \( k - 1 \) different bins. Then,

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

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

\[ \Pr[ \text{every bin } j \text{ has } X_j \leq 1] = \prod_{i=1}^{n-1} \left(1 - \frac{i - 1}{m}\right) \sim \prod_{i=1}^{n-1} e^{-i/m} \]

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

Therefore, \( \Pr[ \text{at least 1 bin } i \text{ has } X_i > 1] \sim 1 - e^{-\frac{n^2}{2m}}. \)
Birthday problem

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 $e^{-\frac{n^2}{2m}} \leq \frac{1}{2} \Rightarrow \frac{n^2}{2m} \leq \ln 2 \sim 0.69$

$\Rightarrow n = \sqrt{2m \ln 2}$. If $m = 365$ then $n = 22.49$.

Therefore, if there are more than 23 students in a class, with probability greater than 1/2, more than 2 students will have the same birthday.
Coupon Collector’s problem

Abraham de Moivre (VIIc.)
How many balls do we need to throw to assure that w.h.p. every bin contains \( \geq 1 \) balls

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

$$\Pr[Y_i = k] = \left(\frac{i - 1}{m}\right)^{k-1} \binom{1 - \frac{i - 1}{m}}{p_i}.$$ 

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

$$\mathbb{E}[Y] = \sum_{i=1}^{m} \mathbb{E}[Y_i] = \sum_{i=1}^{m} \frac{m}{m-i+1} = m \sum_{j=1}^{m} \frac{1}{j} = m(\ln m + o(1)).$$
Coupon Collector’s problem: Concentration

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

- For any bin $j$, define the event $A^r_j$: bin $j$ is empty after the first $r$ throws.
- Notice events $A^r_1, A^r_2, \ldots A^r_m$ are not independent.
- $\Pr[A^r_j] = (1 - \frac{1}{m})^r \sim e^{-r/m}$
- For $r = cm \ln m \Rightarrow \Pr[A^{cm \ln m}_j] \leq e^{-cm \ln m/m} = m^{-c}$.
- Let $W$ be a r.v. counting the number of balls needed to make that every bin has load $\geq 1$.

$$
\Pr[W > cm \lg m] = \Pr[\bigcup_{i=1}^{m} A^{cm \ln m}_j] \leq \sum_{j=1}^{m} \Pr[A^{cm \ln m}_j] \\
\leq \sum_{j=1}^{m} m^{-c} = m^{1-c}.
$$
The previous bound using UB is more tight than the one using Chebyshev or Chernoff on random variable $Y$. (See homework)

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 $\left(\frac{4 \lg n}{\lg \lg n}\right)$, with probability $\leq 1 - \frac{1}{n}$, i.e., w.h.p.

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

To show the above bound we use the following two inequalities:

$$\left(\frac{N}{K}\right)^K \leq \binom{N}{K} \leq \left(\frac{Ne}{K}\right)^K.$$  \hspace{1cm} (1)

Let $N > e$. If $K \geq \frac{2 \ln N}{\ln \ln N}$ then $K^K \geq N$.

$$\left(\frac{N}{K}\right)^K \leq \binom{N}{K} \leq \left(\frac{Ne}{K}\right)^K.$$  \hspace{1cm} (2)
Max-load: Proof Upper Bound

For $1 \leq k \leq n$, $\Pr[X_j \geq k] \leq \binom{n}{k} \frac{1}{n^k} \leq \left(\frac{ne}{k}\right)^k \frac{1}{n^k} \leq \left(\frac{e}{k}\right)^k$.

We want to prove that for $k \geq \frac{2\ln n}{\ln \ln n} \Rightarrow \Pr[X_j \geq \frac{2\ln n}{\ln \ln n}] \leq \frac{1}{n^2}$.

i.e. $\Pr[X_j \geq k] \leq \left(\frac{e}{k}\right)^k \leq \frac{1}{n^2} \Rightarrow \left(\frac{e}{k}\right)^k \geq n^\frac{2}{e}$

Taking $\ln$: $\frac{k}{e} \geq \frac{2\ln(n^{2/e})}{\ln \ln(n^{2/e})} = \frac{4\ln n}{e \ln(\frac{2}{e} \ln n)} \Rightarrow k \geq \frac{4\ln n}{\ln(\frac{2}{e} \ln n)}$

We proved that if $k \geq \frac{4\ln(n)}{\ln(2/e) \ln \ln(n)}$ then $\Pr[X_j \geq k] \leq \frac{1}{n^2}$.

Then, using U-B
$\Pr[\exists i \in [n] \mid X_j \geq k] \leq \sum_{i=1}^{n} \Pr[X_j \geq k] \leq \frac{n}{n^2} = \frac{1}{n}$. 
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 $\Omega\left(\frac{\ln n}{\ln\ln(n)}\right)$ 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 $\Theta\left(\frac{\ln n}{\ln\ln(n)}\right)$.

5. We can obtain sharper bounds for max-load, using strong inequalities (Azuma-Hoeffding) or the Poisson approximation.
1. A difficulty with the exact (binomial) B & B 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.