1 Overview

In this lecture we examined two problems:

- **Coupon Collector**: We are given a set of $n$ different items, and in each timestep we collect a random item from this set. We want to analyze the number of timesteps needed in order to collect all items at least once.

- **Balls and Bins**: We are given $n$ balls and $n$ bins, and we randomly (uniformly) pick a bin to place each of the balls in. Our goal is to analyze certain values of interest for this problem, namely the maximum number of balls in a bin, the concentration of balls among bins and the fraction of bins which are empty, after the balls are distributed.

These problems serve as a starting point for hashing problems.

2 Coupon Collector

**Problem Definition**: We are given a set of $n$ items, and in each timestep we are given a random item from this set. We define $T$ as the number of timesteps required to collect every different item.

**Expected value of $T$**: We define $T_i$ as the number of timesteps required to collect the $(i+1)$-th new item, after we have collected $i$ different items. If we have already collected $i$ items, then the probability that the next item we collect is a new one is $p_i = \frac{n-i}{n}$. This means that the random variables $T_i$ each follow a geometric distribution $T_i \sim \text{Geom} \left( \frac{n-i}{n} \right)$. Thus, we have:

- $\mathbb{E}[T_i] = \frac{1-p_i}{p_i} = \frac{n}{n-i}$
- $\text{Var}[T_i] = \frac{1-p_i}{p_i^2} = \frac{ni}{(n-i)^2}$

We know that $T = \sum_{i=0}^{n-1} T_i$, and that the random variables $T_i$ are independent. Thus, we can derive the following:

$$\mathbb{E}[T] = \sum_{i=0}^{n-1} \mathbb{E}[T_i] = \sum_{i=0}^{n-1} \frac{n}{n-i} = n \sum_{j=1}^{n} \frac{1}{j} = nH_n = \Theta(n \log n) \quad (1)$$

where $H_n = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} = \Theta(\log n)$ is the harmonic number.
**Variance of T:** Similarly, given that the $T_i$ are independent, we have:

$$\text{Var}[T] = \sum_{i=0}^{n-1} \text{Var}[T_i] = \sum_{i=0}^{n-1} \frac{n^2}{(n-i)^2} = n^2 \sum_{i=1}^{n} \frac{1}{j^2} \leq n^2 \sum_{j=1}^{\infty} \frac{1}{j^2} = \frac{n^2 \pi^2}{6}$$  \hspace{1cm} (2)

We also have that $\text{Var}[T] \geq \text{Var}[T_{n-1}] = \frac{n(n-1)}{n} = n^2 - n$. From the above, we have that $\text{Var}[T] = \Theta(n^2)$.

**High probability bounds:** Using Chebyshev’s inequality, we can derive the following, for a given $a$, and setting $\mu = \text{E}[T]$, $\sigma^2 = \text{Var}[T]$:

$$\Pr(|T - \mu| \geq a\sigma) \leq \frac{\text{E}[(T - \mu)^2]}{a^2\sigma^2} = \frac{1}{a^2}$$ \hspace{1cm} (3)

However, this is not a high probability bound $(1 - n^{-c}$, for some constant $c$).

Instead, we can examine the following:

$$\Pr(\text{coupon i is missing after } T \text{ steps}) = \left(1 - \frac{1}{n}\right)^T \leq e^{-\frac{T}{n}}$$ \hspace{1cm} (4)

We can then use the following union bound:

$$\Pr(\text{any coupon is missing after } T \text{ steps}) \leq \sum_{i=1}^{n} \Pr(\text{coupon i is missing after } T \text{ steps}) \leq ne^{-\frac{T}{n}}$$ \hspace{1cm} (5)

By setting the above failure probability to be at most $\delta$, we get:

$$ne^{-\frac{T}{n}} \leq \delta \Rightarrow -\frac{T}{n} \leq \log \frac{\delta}{n} \Rightarrow T \geq n \log \frac{n}{\delta} = \Theta(n \log n)$$ \hspace{1cm} (6)

Thus, if $T$ is above a value scaling as $n \log n$, then all coupons will be collected after $T$ timesteps, with high probability.

### 3 Balls and Bins

**Problem Definition:** We are given $n$ balls along with $n$ bins, and we randomly assign bins to each ball (note that, in a more general setting, the number of balls and bins may differ - here, we will examine this simple case). We set $X_i$ to be the random variable representing the number of balls in bin $i$. Our goal is to examine the load on each bin, as well as the average time it takes to access this structure (retrieve a particular ball from it).

**Maximum value of $X_i$:** The first value of interest we shall examine is $\max_{i=1,\ldots,n} X_i$, the maximum load across all bins. Note that each of the $X_i$ follows a binomial distribution, with probability $p = \frac{1}{n}$, and $n$ trials in total (we have $\frac{1}{n}$ probability when assigning each ball to a bin to pick bin $i$). Thus, $X_i \sim \text{Binom}(n, \frac{1}{n})$. We have the following:

$$\Pr\left(\max_{i=1,\ldots,n} X_i \leq k\right) \leq nPr(X_1 \geq k) = n \binom{n}{k} \left(1 - \frac{1}{n}\right)^{n-k} \leq n \binom{n}{k} \frac{1}{n^k}$$ \hspace{1cm} (7)
Now, we will make use of the inequality \( \binom{n}{k} \leq \left( \frac{en}{k} \right)^k \), and thus obtain from the above:

\[
Pr \left( \max_{i=1,\ldots,n} X_i \right) \leq n \left( \frac{en}{k} \right)^k \frac{1}{n^k} = n \left( \frac{e}{k} \right)^k = \delta
\]  

(8)

Thus, to have probability of failure equal to \( \delta \), we need:

\[
\left( \frac{e}{k} \right)^k = \delta \Rightarrow \log n + k \log \frac{e}{k} = \log \delta \Rightarrow k \log \frac{k}{e} = \log \frac{n}{\delta} = m
\]

(9)

This means that we roughly want \( k \log k \approx \log n = m \). This implies that, for \( \sqrt{m} \leq k \leq m^2 \), we have:

\[
\log k = \Theta(\log m) \Rightarrow m = \Theta(k \log m) \Rightarrow k = \Theta \left( \frac{m}{\log m} \right) \Rightarrow k = \Theta \left( \frac{\log n/\delta}{\log \log n/\delta} \right) = \Theta \left( \frac{\log n}{\log \log n} \right)
\]

(10)

Thus, if we assume \( \delta = n^{-c} \), then in the above we indeed have \( k \log k \approx \log n \), up to constant factors. Thus, the maximum load is \( \Theta \left( \frac{\log n}{\log \log n} \right) \), with high probability.

**Average load over balls:** The next value of interest is the average load over balls, or in other words the value \( E \left[ \sum_{i=1}^{n} X_i^2 \right] \). This value is related to the time we need to retrieve all balls from the bin (to retrieve the balls from bin \( i \), we will need \( X_i \) time to access each of the \( X_i \) balls). We have the following:

\[
E \left[ \sum_{i=1}^{n} X_i^2 \right] = E \left[ \sum_{j=1}^{n} \left( 1 + \# \text{ of balls in the same bin as ball } j \right) \right]
\]

\[
= n + n(n - 1) Pr(\text{balls } j \text{ and } k \text{ fall in the same bin}) \geq 2n - 1
\]

(11)

Thus, our average load over balls scales at least as \( n \).

**Fraction of empty bins:** The final value we examine is the fraction of empty bins, or in other words bins with \( X_i = 0 \). We have the following:

\[
Pr(X_i = 0) = \binom{n}{0} \frac{1}{n^0} \left( 1 - \frac{1}{n} \right)^n = \left( 1 - \frac{1}{n} \right)^n \approx \frac{1}{e} \approx 0.37
\]

(12)

This means that, after the assignment of bins, roughly 37% of the bins are empty. We also note that the same holds for bins with exactly 1 element:

\[
Pr(X_i = 1) = \binom{n}{1} \frac{1}{n^1} \left( 1 - \frac{1}{n} \right)^{n-1} = n \frac{1}{n} \left( 1 - \frac{1}{n} \right)^{n-1} \approx \frac{1}{e} \approx 0.37
\]

(13)

Note: These variables are clearly not independent, but this fact actually helps the concentration bounds to be tighter in our case. Additionally the random variables \( X_i \) are **negatively associated**. In other words, if a subset of the \( X_i \) have a "high value", then all of the other variables must have a "low value". Although there are ways to formalize this notion, these weren’t covered in class.