CS 7280: Data Str and Algs for Scalable Computing Spring 2025

Lecture 7 — February 3, 2025

Prof. Prashant Pandey *Scribe: Bada Kwon*

1 Overview

- This lecture is a primer for **Hashing**
- Basics + Basic Mathematics which we will expand on.

2 Balls and Bins

EXAMPLE: We throw b balls equi-probably and independently into n bins. $(b = n)$

2.1 Applications:

- We can gain insight into **hasing** by studing the balls and bins game
- Hashing is modeled by "randomly" throwing data into hash table buckets.
- Another application is **load balancing**. Bins can be seen as **servers**, and balls can be seen as **clients**

2.2 Questions to answer for Balls and Bins:

1. Expected number of balls in a bin?
2. Expected number of balls in the fullest bin?
3. Expected number of balls thrown before getting a collision?
4. There are more... But These three will be the focus.

What happens when we change "Expected number" with "high probability"
• Gives us a result we can use in the real world, becuase it gives us the "higher probability" bound for that event.

3 Review of basic probability

3.1 Definition 1: Probability Sample Space

A probability sample space is defined as (S, P) where:

- $S = \{s_1, s_2, ..., s_n\}$ is the set of all possible outcomes.
- $P : S \to [0, 1]$ assigns probabilities to outcomes.
- The sum of probabilities satisfies $\sum P(s_i) = 1$.

3.2 Definition 2: Event

An **event** is a subset of outcomes from the sample space S.

Steps for solving event probability problems:

1. Find the sample space.
2. Define events of interest.
3. Determine outcome probabilities.
4. Determine event probabilities.

3.3 Definition 3: Random Variable

A **random variable** is a function:

$$f : S \to \mathbb{R}^+ \tag{1}$$

Example assignments:

- If $S \to$ Heads, then $f = 1$.
- If $S \to$ Tails, then $f = 0$.

3.4 Definition 4: Expected Value

The **expected value** of a random variable f is:

$$E[f] = \sum P(s_i) \cdot f(s_i) \tag{2}$$

3.5 Definition 5: Linearity of Expectation

For any two functions f and g:

$$E[f + g] = E[f] + E[g] \tag{3}$$

3.6 Definition 6: Conditional Probability

The conditional probability of A given B is:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{4}$$

3.7 Definition 7: Independence

Two events A and B are **independent** if:

$$P(A \cap B) = P(A)P(B) \tag{5}$$

3.8 Definition 8: Mutual Independence

Events $E_1, E_2, ..., E_n$ are **mutually independent** if:

$$P(E_1 \cap E_2 \cap ... \cap E_n) = P(E_1)P(E_2)...P(E_n) \tag{6}$$

Every variable is independent of any combination of other variables in the set.

3.9 Definition 9: Pairwise Independence

Events $E_1, E_2, ..., E_n$ are **pairwise independent** if:

$$P(E_i \cap E_j) = P(E_i)P(E_j) \quad \text{for all distinct } i, j \tag{7}$$

Every pair of variables within a set are independent of each other, but it doesn't necessarily mean that any combination of three or more variables are independent.
Pairwise independence is a weaker condition than mutual independence.

3.10 Definition 10: High Probability

VERY VERY IMPORTANT.

Let E_n be an event on problem size n. We say that E_n occurs **with high probability** if:

$$P[E_n] = 1 - \frac{1}{n^c}, \quad \text{for some constant } c \geq 1 \tag{8}$$

$$\lim_{n \to \infty} P[E_n] = 1 \tag{9}$$

4 Solving the Questions

4.1 Q1: Expected Number of Balls in Bin 1

Given that there are a total of n balls and n bins:

- Each ball is placed into a bin independently and uniformly at random.
- The probability of any specific ball landing in bin 1 is $\frac{1}{n}$.
- Let X_i be an indicator random variable such that:

$$X_i = \begin{cases} 1, & \text{if ball } i \text{ lands in bin 1} \\ 0, & \text{otherwise} \end{cases}$$

- Then, the total number of balls in bin 1 is:

$$X = \sum_{i=1}^{n} X_i$$

- Since $E[X_i] = P(X_i = 1) = \frac{1}{n}$:
- By linearity of expectation:

$$E[X] = \sum_{i=1}^{n} E[X_i] \ = \frac{1}{n} + \frac{1}{n} + ... + \frac{1}{n}$$

$$E[X] = n \cdot \frac{1}{n} = 1$$

4.2 Q2: Number of Balls in the Fullest Bin With High Probability

Given n balls and n bins, we aim to determine the number of balls in the fullest bin **with high probability**.

One of the most critical proofs for Hashing

- What is the event of interest???

Proof:

We start by computing the probability that bin 1 has exactly l balls:

$$P[\text{bin 1 has } l \text{ balls}] = \binom{n}{l} \left(\frac{1}{n}\right)^l \left(1 - \frac{1}{n}\right)^{n-l}$$

where:

- $\binom{n}{l}$ represents the number of ways to choose l balls from n balls.
- $\left(\frac{1}{n}\right)^l$ is the probability that these l balls land in bin 1.
- $\left(1 - \frac{1}{n}\right)^{n-l}$ is the probability that the remaining $n - l$ balls do not land in bin 1.

Now we find the probability that bin 1 has more than l balls

$$P[\text{bin 1 has more than } l \text{ balls}] \leq \binom{n}{l} \left(\frac{1}{n}\right)^l$$

Death Bed Forumlas:
Always remember these!!!

- $$(\frac{y}{x})^x \le \binom{y}{x} \le (\frac{ey}{x})^x$$

- $$P(A \cup B) = P(A) + P(B) - P(A \cap B) \le P(A) + P(B)$$

Back to solving Q2...

$$P[\text{bin 1 has more than } l \text{ balls}] \le \binom{n}{l}\left(\frac{1}{n}\right)^l$$

$$\le (\frac{en}{ln})^l$$

$$\le (\frac{e}{l})^l$$

Intuition:
Let's say $l = clgn$

$$P[\text{any bin has} \ge clgn \text{ balls}] \le n(\frac{e}{clgn})^{clgn}$$

$$\text{NOTE: plotting } \frac{e}{clgn} \text{ is bounded by } \frac{1}{2}$$

$$\le n(\frac{1}{2})^{clgn}$$

$$\le n * n^{-c}$$

$$\le n^{1-c}$$

This is still a loose bound!!!

We want to get to... "with high probability":

$$P[E_n] = 1 - \frac{1}{n^c}, \quad \text{for some constant } c \ge 1$$

5 Recap

1. Find the "fullest" bin, meaning the maximum number of balls in any bin, denoted as l.
2. Bound any bin's chance than more than l balls → find the chance for 1 bin

3. Linearity of Expectation gives us n^{1-c} (with loose bound $l = clgn$)

$$P[\text{any bin has at least } l \text{ balls}] \leq n^{1-c}$$

- hidden trials to find lower bound of l
- TRY IT OUT: Solve for $l = \frac{clgn}{lglgn}$...