Balls and Bins

m balls, n bins. $\forall i \in [m]$, throw ball i to random bin j. $(j \in [n])$
independently.

How are balls distributed across n bins?

One of the most natural random processes
- Coupon Collector
- "Hashing"

Questions.
- For each m, # balls in largest/smallest bin?
- When does every bin become nonempty?

$m = 0$. $\qquad m = n \cdot H_n \qquad m = C \cdot n \log n$

$1 + \frac{1}{2} + \cdots + \frac{1}{n} = \ln n \pm O(1)$ $\qquad$ (C large constant)

When do two items occupy the same bin? (Birthday Paradox)

Let $S :=$ # of (unordered) pairs (j,k) of balls in the same bin.

$\mathbb{E}[S] = \binom{m}{2} \cdot 1/n = \frac{m(m-1)}{2n}$ Markov

So, if $m = \alpha\sqrt{n}$, $\Pr[S \geq 1] \leq \frac{m(m-1)}{2n} = O(\alpha^2)$.

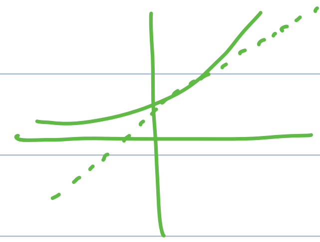
More precisely, for $m \leq n$, let $p_{m,n} = \Pr[\text{all bins have} \leq 1 \text{ ball}]$.

$p_{m,n} = 1 \cdot (1 - 1/n) \cdot (1 - 2/n) \cdot \dots \cdot (1 - {}^{m-1}/n)$. $1 + x \leq e^x$ for "all x"

$\leq e^{-(1/n + 2/n + \dots + {}^{m-1}/n)}$

$= e^{-\binom{m-1}{2}/n}$.

So if $m = \alpha\sqrt{n}$ $p_{m,n} \leq e^{-\Omega(\alpha^2)}$.



$e^x = 1 + x + x^2/2 + x^3/6 + \dots$

$p_{m,n} = 1 \cdot (1 - 1/n) \cdot (1 - 2/n) \cdot \dots (1 - {}^{m-1}/n)$

$\geq e^{(-1/n + 1/n^2 - 2/n + (2/n)^2 - \dots - {}^{m-1}/n + ({}^{m-1}/n)^2)}$ $1 + x + x^2 \geq e^x$ for $x \in [-1, +1]$

$= e^{-\binom{m-1}{2}/n + O(m^3/n^2)}$

So if $m = \alpha\sqrt{n}$, $p_{m,n} \geq e^{O(-\alpha^2 + \alpha^3/\sqrt{n})}$

When does every bin become nonempty? :)

Let S_i be the # balls when i bins are nonempty for the first time, and $T_i = S_{i+1} - S_i$ (let $S_0 = 0$)

$S_n = T_0 + \cdots + T_{i-1}$. Each T_i is independent!

Fix i. T_i is "geometrically distributed with parameter $p_i = \frac{n-i}{n}$".

$$T_i = \begin{cases} 1 & \text{w.p. } p_i \\ 2 & \text{w.p. } (1-p_i)p_i \\ 3 & \text{w.p. } (1-p_i)^2 p_i \\ \vdots & \\ t & \text{w.p. } (1-p_i)^{t-1} p_i. \end{cases}$$

$\mathbb{E}[T_i] = 1/p_i$, $\sigma^2[T_i] = (1-p_i)/p_i^2$.

Then, $$\mathbb{E}[S_n] = \sum_{i=0}^{n-1} \frac{1}{p_i} = \sum_{i=0}^{n-1} \frac{n}{n-i} = n\left(\frac{1}{n} + \cdots + 1\right) = nH_n!$$

$$\sigma^2(S_n) = \sum_{i=0}^{n-1} \frac{(1-p_i)}{p_i^2} = \sum_{i=0}^{n-1} \frac{in}{(n-i)^2} \overset{j=n-i}{=} \sum_{j=1}^{n} \frac{n(n-j)}{j^2} = n^2 \underbrace{\left(1 + \frac{1}{2^2} + \cdots + \frac{1}{n^2}\right)}_{\leq \pi^2/6 \leq 2.} - nH_n.$$

$$\Rightarrow \sigma(S_n) \leq \sqrt{2}\, n.$$

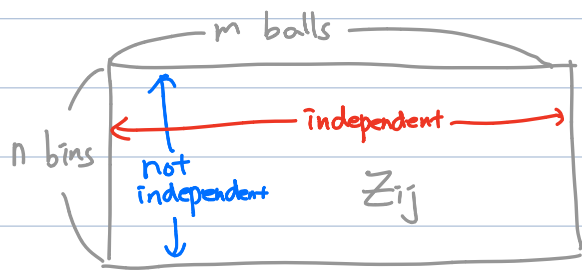
So, $\Pr[|S_n - nH_n| \geq kn] \leq \Pr\left[|S_n - nH_n| \geq \frac{k}{\sqrt{2}} \cdot \sigma(S_n)\right] \leq 2/k^2$.

(much stronger bound possible)

When $m = C \cdot n \log n$ for large C,

$\forall i \in [m], j \in [n]$, $Z_{ij} = 1$ if ball i is in bin j, 0 o.w.

$\forall j \in [n]$, let $X_j = \sum_{i=1}^{m} Z_{ij}$



Another version of Chernoff:
if $Y = Y_1 + \cdots + Y_k$ is the **sum** of i.i.d. $\{0,1\}$-valued rv's,

$$\Pr[|Y - \mathbb{E}[Y]| \geq \varepsilon \mathbb{E}[Y]] \leq 2e^{-\varepsilon^2 \mathbb{E}[Y]/3}$$

For each $j \in [n]$, $X_j = \sum_{i=1}^{m} Z_{ij}$ is sum of i.i.d. 0-1 random variables.

Let $X = X_j / m$ (average of Z_{ij}), and use Chernoff

$(\mathbb{E}[X] = 1/n)$

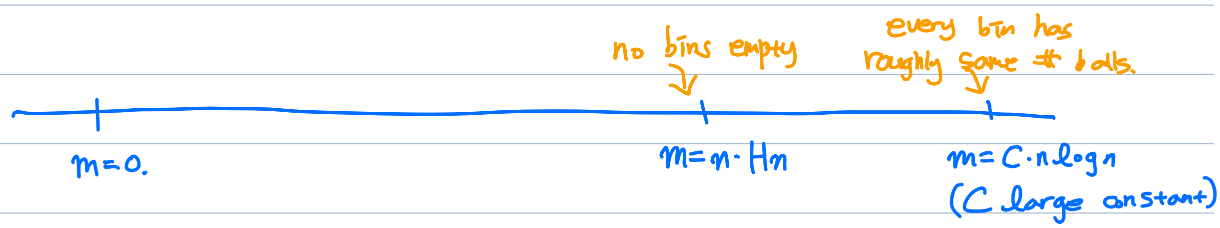
$$\Pr[|X - \mathbb{E}[X]| \geq \varepsilon \mathbb{E}[X]] \leq 2e^{-\varepsilon^2 \overbrace{k\mathbb{E}[X]}^{= m/n}/3}$$

Using $\varepsilon = 1/2$, $\Pr[|X_j/m - 1/n| \geq 1/2n] \leq 2e^{-m/12n} \leq 2/n^2$ if $C \geq 24$.

By union bounding over all $j \in [n]$

$$\Pr[|X_j - m/n| \geq m/2n \text{ for some } j] \leq 2/n.$$

W.p. $\geq 1 - 2/n$, all bins have $[m/2n, 3m/2n]$ balls!



Verifying Matrix Product

This lecture: given A and B, is $A = B$?
- But A and B are given in "complex forms", so deterministic checking may be inefficient.
- Randomness and algebraic techniques help!

Verifying Matrix Product

Input: $A, B, C \in \mathbb{R}^{n \times n}$

Output: YES if $AB = C$, NO o.w.

Deterministic algo: Compute AB and check $AB = C$.
$O(n^{2.37\ldots})$ time (very complicated)

Randomized algo.

$\forall i \in [n]$, sample d_i from $\{0,1\}$

Sample d from $\{0,1\}^n$

Compute $ABd = A(Bd)$

Compute Cd

YES if $ABd = Cd$, NO o.w.

3 matrix-vector multiplication: Running time $O(n^2)$.

If $AB = C$, then $ABd = Cd$ $\forall d$, so YES w.p. 1

(with probability)

If $AB \neq C$, let $(AB)_i$ and C_i be the i^{th} column of AB and C respectively and assume $(AB)_{i^*} \neq C_{i^*}$ for some $i^* \in [n]$.

Then given any fixed sampled values of $d_1, \dots, d_{i^*-1}, d_{i^*+1}, \dots d_n$,

$$\Pr_{d_{i^*}}[ABd = Cd] = \Pr_{d_{i^*}}\left[\sum_{j=1}^{n} (AB)_j d_j = \sum_{j=1}^{n} C_j d_j\right]$$

consider only d_{i^*} as random

$$= \Pr_{d_{i^*}}\left[\sum_{j=1}^{n} (AB-C)_j d_j = 0\right]$$

$$= \Pr_{d_{i^*}}\Big[\underbrace{\sum_{j \neq i^*} (AB-C)_j d_j}_{\text{fixed vector } e} = -(AB-C)_{i^*} d_{i^*}\Big]$$

$$= \begin{cases} \frac{1}{2} & \text{if } e = 0 \text{ or } -(AB-C)_{i^*} \\ 0 & \end{cases}$$

$\therefore$ If $AB \neq C$, will output NO w.p. $\geq 1/2$!

("error probability" $\leq 1/2$)

Repeating t times will make error probability $\leq 1/2^t$

$$\begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \stackrel{?}{=} \begin{bmatrix} 4 & 7 \\ 7 & \overset{10}{11} \end{bmatrix}$$

$$\begin{bmatrix} 5 & 7 \\ 8 & 10 \end{bmatrix} \begin{bmatrix} d_1 \\ d_2 \end{bmatrix} = \begin{bmatrix} 4 & 7 \\ 7 & \substack{10 \\ 11} \end{bmatrix} \begin{bmatrix} d_1 \\ d_2 \end{bmatrix}$$

$$\begin{bmatrix} 5 \\ 8 \end{bmatrix} d_1 + \begin{bmatrix} 7 \\ 10 \end{bmatrix} d_2 = \begin{bmatrix} 4 \\ 7 \end{bmatrix} d_1 + \begin{bmatrix} 7 \\ \substack{10 \\ 11} \end{bmatrix} d_2$$

$$\begin{bmatrix} 1 \\ 1 \end{bmatrix} d_1 = \begin{bmatrix} 0 \\ \substack{0 \\ 1} \end{bmatrix} d_2$$