Randomized Algorithms

Coupon collector problem: n different coupons, sampled with replacement from uniform distribution.

Q: How many coupons before one of each kind shows up?

Define $X_i = \#$ of rounds between $i$th and $(i+1)$st coupon

$X_i \sim \text{Geo} \left( \frac{n-i}{n} \right) = \text{Geo} \left( 1 - \frac{i}{n} \right)$

$E(X_i) = \frac{n}{n-i}$.

$\Rightarrow E \left( \sum_{i=0}^{n-1} X_i \right) = \sum_{i=0}^{n-1} E[X_i] = \sum_{i=0}^{n-1} \frac{n-i}{n} = n \log n = \Theta(n \log n)$.

Linearity of expectation: For any set of random variables $X_1, \ldots, X_n$ (not necessarily independent!!)

$E \left( \sum_{i=1}^{n} X_i \right) = \sum_{i=1}^{n} E[X_i]$

Balls and bins

Suppose we throw n balls uniformly at random into n bins.

Let $X_i$: # of balls in the $i$th bin

$E[X_i] = 1 \quad \forall \ i$ — not very interesting

Let $X = \max_i X_i$

What is $E[X]$?

For any $i$, $P[X_i \geq k] \leq \left( \frac{n}{i} \right) \frac{1}{n} \leq \left( \frac{ne}{k} \right) \frac{1}{n} = \left( \frac{x}{k} \right)$

$P[X_i \geq \frac{e \ln n}{\ln \ln n}] \leq \left( \ln \ln n \right)^{-c} \text{ for some } c > 0$.
\[ \Pr \left[ X \geq \frac{e^{\ln n}}{\ln \ln n} \right] \leq \left( \frac{\ln \ln n}{\ln n} \right)^{\ln \ln n} \leq \frac{1}{n} \]

**Union Bound**
\[ \Pr \left[ \bigcup_i E_i \right] \leq \sum_i \Pr \left[ E_i \right] \]

Using union bound,
\[ \Pr \left( X \geq \frac{e^{\ln n}}{\ln \ln n} \right) \leq \frac{1}{n} \]

\[ \mathbb{E}[X] \leq \left( 1 - \frac{1}{n} \right) \frac{e^{\ln n}}{\ln \ln n} + \frac{1}{n} \cdot n = O\left( \frac{\ln n}{\ln \ln n} \right) \] (also called a tail bound)

We need a tail bound to obtain a bound on the expectation. Often we want the reverse: use expectation to obtain a tail bound.

**Markov's Inequality**: \[ \Pr \left[ X \geq t \right] \leq \mathbb{E}[X] / t \text{ for } X \geq 0. \]

**Proof**: \[ \mathbb{E}[X] \geq \Pr \left[ X \geq t \right] \cdot t + \Pr \left[ X < t \right] \cdot 0 \]

Often too weak by itself (e.g. in balls in bins, gives probability bound of \( \ln n / \ln \ln n \) which is too weak for union bound) but useful for proving stronger inequalities.

**Chebyshev's Inequality**
\[ \Pr \left[ |X - \mu| \geq t \sigma \right] \leq \frac{1}{t^2} \] (Note: \( X \) need not be non-negative)

**Proof**: Using Markov,
\[ \Pr \left[ (X - \mu)^2 \geq t^2 \sigma^2 \right] \leq \frac{\mathbb{E}[(X - \mu)^2]}{t^2 \sigma^2} = \frac{\sigma^2}{t^2 \sigma^2} = \frac{1}{t^2} \]
Chernoff Bounds

For independent $X_1, \ldots, X_n$ where $X_i = 1$ w.p. $p_i$ and 0 otherwise,

$$\Pr\left[ X = \sum_{i=1}^{n} X_i > (1+\varepsilon)\mu \right] < \left( \frac{e^\varepsilon}{(1+\varepsilon)^{1+\varepsilon}} \right)^\mu$$

$$\Pr\left[ X < (1-\varepsilon)\mu \right] < \left( \frac{e^{-\varepsilon}}{(1-\varepsilon)^{1-\varepsilon}} \right)^\mu$$

Proof: Let us prove only positive deviation.

$$\Pr\left[ X > (1+\varepsilon)\mu \right] = \Pr\left[ e^{tx} > \exp(t(1+\varepsilon)\mu) \right] \leq \frac{\mathbb{E}[e^{tx}]}{\exp(t(1+\varepsilon)\mu)}$$

Monotonically increasing function of $t$ for $t > 0$

$$\prod_{i=1}^{n} \frac{e^{tx_i}}{\exp(t(1+\varepsilon)\mu)} = \frac{\prod_{i=1}^{n} e^{tx_i} \cdot p_i + (1-p_i)}{\exp(t(1+\varepsilon)\mu)} \leq \exp(\ln(p_i) (e^t - 1))$$

Simpler (more useful bounds):

$$\varepsilon \in (0, 1]: \begin{cases} 
\Pr\left[ X > (1+\varepsilon)\mu \right] \leq \exp\left(-\frac{\varepsilon^2 \mu}{3}\right) \\
\Pr\left[ X < (1-\varepsilon)\mu \right] \leq \exp\left(-\frac{\varepsilon^2 \mu}{2}\right)
\end{cases}$$

Generic Counting via Sampling
Suppose you have a universe \( U \) and subset \( S \subseteq U \).

Your goal is to estimate \(|S|\), given a uniform sampling procedure for \( U \).

Algorithm: \( X_i = \begin{cases} 1 & \text{if sample in } S \\ 0 & \text{o.w.} \end{cases} \)

Output: \( \left( \frac{\sum_{i=1}^{N} X_i}{N} \right) |U| = \tilde{N}_S \leftarrow \text{unbiased estimator} \quad E[\tilde{N}_S] = |S| \)

Q: How many samples \( N \) do you need?

Suppose \( |S|/|U| = \theta \)

\[ \Pr \left[ |\tilde{N}_S - |S|| > \varepsilon |S| \right] = \Pr \left[ |\sum_{i=1}^{N} X_i - \theta N| > \varepsilon \theta N \right] \leq e^{-\frac{2\varepsilon^2 N \theta^2}{3}} \]

by Chernoff

For \( O(1) \) error probability, \( N \geq \frac{3}{\varepsilon^2 \theta} = \frac{3|U|}{\varepsilon^2 |S|} \)

For \( O(1/n^d) \) error probability, repeat the \( n \) times (BOOSTING)

large of \( (s) \) small compared to \( |U| \)