1: Balls and Bins

$$\boxed{Markov's\ Inequality:\ Pr(X \geq a) \leq \frac{E[x]}{a}} \tag{1}$$

(a) Given n bins and $4n \log n$ balls, we want to prove that the probability that there exists an empty bin is $< 1/n$.

Proof We will prove this using Markov's Inequality. Let X be a random variable representing the number of empty bins. We will show that the probability that at least one bin is empty is $< 1/n$.

$$Pr(X \geq 1) \leq E[X]$$

From class, we saw the expectation of X is equivalent to the summation of each of the bins. Let y_i represent each bin where a 1 represents an empty bin and a 0 otherwise.

$$Pr(X \geq 1) \leq E[X] = E[\sum_{i=1}^{n} y_i]$$

The probability a bin is empty is given by $(1 - 1/n)^m$ where m is the number of balls. That is, there are $1 - 1/n$ other bins each ball can be placed in.

$$Pr(X \geq 1) \leq n * (1 - \frac{1}{n})^{4nlogn}$$

From Bernoulli's inequality we know that $1 + y \leq e^y$ for all for all y. Letting $y = \frac{-1}{n}$ will allow us to substitute $1 + y$ for e^y. This is valid as $1 + y \leq e^y$ so Markov's Inequality still holds.

$$Pr(X \geq 1) \leq n * (1 + y)^{4nlogn}$$
$$Pr(X \geq 1) \leq n * (e^y)^{4nlogn}$$
$$Pr(X \geq 1) \leq n * (e^{-1/n})^{4nlogn}$$

Applying exponent and natural log identities.

$$Pr(X \geq 1) \leq n * e^{-4logn}$$
$$Pr(X \geq 1) \leq \frac{1}{n^3} < \frac{1}{n}$$

Thus proving the probability that at least one bin is empty is $< 1/n$.

(b.a) When $m = \frac{1}{2}nlogn$, the logic follows as in part a.

$$Pr(X \geq 1) \leq E[X] = E[\sum_{i=1}^{n} y_i] = n * (1 - \frac{1}{n})^{\frac{1}{2}nlogn}$$

Where y_i is a bin and 1 represents an empty bin and 0 otherwise. Making the similar substitution and applying similar identities as part a.

$$Pr(X \geq 1) \leq n * (e^{-1/n})^{\frac{1}{2}nlogn}$$
$$Pr(X \geq 1) \leq n * \frac{1}{\sqrt{n}}$$
$$Pr(X \geq 1) \leq \sqrt{n}$$

Given $\frac{1}{2}nlogn$ balls, we can say that the probability that at least one bin is empty is $\leq \sqrt{n}$.

$\boxed{Pr(X \geq 1) \leq \sqrt{n}}$

(b.b) When $m = 100nlogn$, a similar logic follows.

$$Pr(X \geq 1) \leq E[X] = E[\sum_{i=1}^{n} y_i] = n * (1 - \frac{1}{n})^{100nlogn}$$
$$Pr(X \geq 1) \leq n * (e^{-1/n})^{100nlogn}$$
$$Pr(X \geq 1) \leq n * \frac{1}{n^{100}}$$
$$Pr(X \geq 1) \leq \frac{1}{n^{99}}$$

Given $100nlogn$ balls, we can say that the probability that at least one bin is empty is $\leq \frac{1}{n^{99}}$.

$\boxed{Pr(X \geq 1) \leq \frac{1}{n^{99}}}$

(c) Given n bins and n balls we want to bound the probability that 90% of the bins are empty. Using Markov's Inequality we will derive such a bound.

$$Pr(X \geq a) \leq \frac{E[X]}{a}$$
$$Pr(X \geq 0.9n) \leq \frac{E[X]}{0.9n}$$

Where X is a random variable representing the number of empty bins. From class we the expectation of X is equivalent to the summation of each of the bins. Let y_i represent each bin where a 1 represents an empty bin and a 0 otherwise.

$$Pr(X \geq 0.9n) \leq \frac{E[\sum_{i=1}^{n} y_i]}{0.9n}$$
$$Pr(X \geq 0.9n) \leq \frac{n * (1 - \frac{1}{n})^n}{0.9n}$$
$$Pr(X \geq 0.9n) \leq \frac{n * (e^{\frac{-1}{n}})^n}{0.9n}$$
$$Pr(X \geq 0.9n) \leq \frac{\frac{1}{e}}{0.9} = \frac{1}{0.9e}$$

Given n balls and n bins, we can say that the probability that 90% of the bins are empty is $\leq \frac{1}{0.9e}$.

$\boxed{Pr(X \geq 0.9n) \leq \frac{1}{0.9e}}$

(d) From the problem, we know that for any distinct indices $j_1, j_2, ..., j_k$, we have

$$\textbf{Inequality One}: Pr[X_{j1} = 1 | X_{j2} = X_{j3} = ... = X_{jk} = 1] \leq Pr[X_{j1} = 1]$$

From Baye's Rule we have:

$$Pr[A = 1 \;\&\; B = 1] = Pr[A|B] * P[B]$$

We want to prove that the probability that 90% of the bins are empty is at most $(0.9)^n$.

As a first step, we will first use the given inequality and Baye's rule to prove:

$$\textbf{Inequality Two}: Pr[X_{j1} = X_{j2} = X_{j3} = ... = X_{jk} = 1] \leq e^{-k}$$

That is, the probability that a given k bins are all empty is $\leq e^{-k}$.

First, we will use Baye's Rule. Assigning two new random variables A and B. Where A is $X_{j1} = 1$ and B is $X_{j2} = X_{j3} = ... = X_{jk}$. The goal is to transform the LHS of *Inequality Two* to be identical to the LHS of *Inequality One*.

$$Baye's\ Rule: \; Pr[A = 1 \;\&\; B = 1] = Pr[A|B] * P[B]$$

$$What\ we\ start\ with: \; Pr[X_{j1} = X_{j2} = X_{j3} = ... = X_{jk} = 1]$$

$$Applying\ Baye's\ Rule: \; Pr[X_{j1} = 1 | X_{j2} = .. = X_{jk}] * Pr[X_{j2} = ... = X_{jk}]$$

Now Baye's Rule is applied again to $Pr[X_{j2} = ... = X_{jk}]$ where A is $X_{j2} = 1$ and B is $X_{j3} = ... = X_{jk} = 1$:

$$What\ we\ start\ with: \; Pr[X_{j2} = X_{j3} = X_{j4} = ... = X_{jk} = 1]$$

$$Applying\ Baye's\ Rule: \; Pr[X_{j2} = 1 | X_{j3} = .. = X_{jk}] * Pr[X_{j2} = ... = X_{jk}]$$

This pattern of applying Baye's rule to the $P[B]$ portion is repeated k times. This produces a product of the form:

$$\prod_{i=1}^{k} Pr[X_{ji} = 1 | X_{j(i+1)} = ... = X_{ji} = 1]$$

This is identical in form of the LHS of *Inequality One* if we take a product over both sides of *Inequality One*:

$$\prod_{i=1}^{k} Pr[X_{ji} = 1 | X_{j(i+1)} = ... = X_{ji} = 1] \leq \prod_{i=1}^{k} Pr[X_{ji}]$$

Thus, we need to analyze the RHS of the above inequality to prove the bounds for *Inequality Two*. We saw in class that the probability that bin j is empty is $(1-\frac{1}{n})^n \leq \frac{1}{e}$ where n is the number of balls. Using this, we can take the product of $\frac{1}{e}$ k times.

$$\prod_{i=1}^{k} Pr[X_{ji}] \leq \prod_{i=1}^{k} \frac{1}{e} = e^{-k}$$

Thus proving *Inequality Two*.

Using *Inequality Two*, we can get a bound on the probability that 90% of the bins are empty. Let $k = 0.9n$.

$$Pr[X_{j1} = X_{j2} = ... = X_{jk} = 1] \leq e^{-0.9n}$$

Returning to the original proof, we want to prove the probability that 90% of the bins are empty is bounded by 0.9^n. We have proven the bound $e^{-0.9n}$. We know that n is a positive integer as it's the number of bins. For all positive integers $e^{-0.9n} \leq 0.9^n$ holds. Thus we have proven a tighter bound and in the process proven the original bound.

2: Estimating the Mean and Median

As by the suggestion of the TA Michael Matteny I will be using Hoeffding's Inequality to provide bounds for the following questions rather than the suggested Chernoff bounds. They are related and provide similar strength in bounds as they both provide bounds for **independent** random variables. This Hoeffding's Inequality stronger than Markov's or Chebyshev's inequalities.

$$\boxed{Hoeffding's\ Inequality:\ P(|X - E[X]| \geq t) \leq 2exp(\frac{-2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2})} \tag{2}$$

(a) This variation of Hoeffding's Inequality is used when we know that X_i's are strictly in the intervals $[a_i, b_i]$. For our purposes a_i is -1 and b_i is 1. Let X be the sample mean after j random indices are sampled or $\hat{\mu}$. The expected value of X is the true mean or μ.

Hoeffding's Inequality tells us the probability that the difference between X and $E[X]$ will be $\geq t$ is less than the RHS of the inequality. That is, if t is ϵ we want the RHS to be $\leq \delta$. This will give us the number of required samples to ensure that $|\hat{\mu} - \mu| \leq \epsilon$ with probability $1 - \delta$, where n is the number of samples.

$$P(|\hat{\mu} - \mu| \geq \epsilon) \leq 2exp(\frac{-2n^2\epsilon^2}{\sum_{i=1}^{n}(1+1)^2}) \leq \delta$$

$$exp(\frac{-2n^2\epsilon^2}{4n}) \leq \frac{\delta}{2}$$

$$\frac{-n\epsilon^2}{2} \leq \ln(\frac{\delta}{2})$$

$$n\epsilon^2 \geq ln(\frac{4}{\delta^2})$$

$$\boxed{n \geq ln(\frac{4}{\delta^2}) * \frac{1}{\epsilon^2}} \tag{3}$$

Given a δ and ϵ, the derived equation produces the n number of indices we must sample to satisfy $|\hat{\mu} - \mu| \leq \epsilon$ with probability $1 - \delta$.

(b) No. If we were to sample without replacement, the sampled a_j's are no longer independent events. They become **dependent** on what has already been sampled as the sample space is changing with each sample. Hoeffding's Inequality is only applicable for **independent** random variables.

(c) Given the value of each a_i has the constraint $a_i \in [-M, M]$, we can still use Hoeffding's Inequality to derive the require number of samples. The setup will be similar to part a, with the exception that now $a = -M$ and $b = M$. Recall that the generalization of Hoeffding's Inequality is used when we know that X_i's are strictly bounded by the intervals $[a_i, b_i]$.

$$P(|\hat{\mu} - \mu| \geq \epsilon) \leq 2exp(\frac{-2n^2\epsilon^2}{\sum_{i=1}^{n}(M+M)^2}) \leq \delta$$

$$exp(\frac{-n\epsilon^2}{2M^2}) \leq \frac{\delta}{2}$$

$$\frac{-n\epsilon^2}{2M^2} \leq ln(\frac{\delta}{2})$$

$$\frac{n\epsilon^2}{M^2} \geq ln(\frac{4}{\delta^2})$$

$$\boxed{n \geq ln(\frac{4}{\delta^2}) * \frac{M^2}{\epsilon^2}} \tag{4}$$

Given a δ, ϵ, and a M, the derived equation produces the n number of indices we must sample to satisfy $|\hat{\mu} - \mu| \leq \epsilon$ with probability $1 - \delta$.

(d) Proof by Counter-Example:

3: Quick-sort with Optimal Comparisons

(a) The goal is to prove the following statement about the probability the k'th smallest element in A is chosen as the pivot.

$$p_k = \frac{\binom{k-1}{m}\binom{n-k}{m}}{\binom{n}{2m+1}}$$

We sample M random elements from A (without replacement), and pick the median of these entries as a pivot, where $M = 2m + 1$ for an integer $m \geq 1$. Meaning, in our sample there will be m numbers less than or equal to the chosen pivot and m numbers greater than or equal to the pivot.

For the chosen pivot to be the k'th smallest element in A, there are $k-1$ possible numbers to pick from that are smaller than k and $n-k$ numbers to pick that are larger than k. That is, m numbers in the range $k-1$ and m numbers in the range $n-k$ must be chosen. This can be interpreted as $k-1$ choose m and $n-k$ choose m. When both these events happen, the k'th smallest element has been chosen as the pivot (so we take their product).

The space of all possible choices is n choose $2m+1$. We have n items in A[] and we are selecting $2m+1$ items. Thus we have proven the equation.

$$p_k = \frac{\binom{k-1}{m}\binom{n-k}{m}}{\binom{n}{2m+1}}$$

(b) Let X be a random variable for the number of comparisons for a given array size. Using the p_k from part a we can construct a recursive formula for the expected number of comparisons on an array of size n.

$$E[X] = \sum_{i=1}^{n} p_i * x_i$$

$$E[X_n] = \sum_{k=1}^{n} p_k(E[X_{k-1}] + E[X_{n-k}])$$

The expected number of comparisons on an array of size n is recursively dependent on the expected number of comparisons required on the two sub-arrays created when we split on the pivot in the quick-sort algorithm.

4: Randomized Min-Cut

(a) The min-cut of an undirected graph $G = (V, E)$ is a partition of the nodes into two disjoint sets V_1 and V_2 s.t. the number of edges between V_1 and V_2 is minimized.

Let E' be the set of edges that separate V_1 and V_2. Say we collapse an edge that is not in E', meaning we connect it's end points into a supernode which gives a new graph with one less vertex. Let G after this collapse be G'.

As a result of this collapse, the degree of all $v \in V$ either increase or stay the same. This is because we connect the removed edges endpoints with parallel edges.

As a result of the degree of all $v \in V$ in G' not decreasing, the min-cut of G' will be equal to that in G. The only way to find a smaller min-cut than E' is if the degree of a vertex reduced as a result of the collapse.

(b) We know that the total degree of all the vertices in G is $2|E|$ from the degree sum formula.

$$\sum_{v \in V} deg(v) = 2|E|$$

Let $n = |V|$. We can make a statement about the average degree of a vertex in G.

(c)

(d)

5: Valiant-Vazirani Lemma