1: Balls and Bins

(a) The probability of a bucket being empty with n buckets and m balls is $(1-\frac{1}{n})^m$. For the case when $m = 4n\log n$,

$$= (1-\tfrac{1}{n})^{4n\log n} = (1-\tfrac{1}{n})^{n\log n^4}$$
$$= (\tfrac{1}{e})^{\log n^4} = \tfrac{1}{n^4} < \tfrac{1}{n}$$

(b) Using the above steps,

1. Probability of a bin being empty when $m = \frac{1}{2}n\log n = (\frac{1}{\sqrt{n}})$
2. Probability of a bin being empty when $m = 100n\log n = (\frac{1}{n^{100}})$

From these values we can say that as value of m increases linearly the probability of a bin being empty decrease exponentially.

(c) The expected number of bins as we see in the class is n/e (when $m = n$). Let X be the event that a bi is empty then, $E[X] = n/e$. The value that we nee is $Pr[X \geq 0.9n]$ or $Pr[X \geq \frac{9e}{10}\frac{n}{e}]$. UsingMarkov's inequality this probability is less than equal to $\frac{10}{9e}$. Therefore, $Pr[X \geq 0.9n] \leq \frac{10}{9e}$.

(d) We know that the conditional probability $P(A|B) = \frac{P(A\cap B)}{P(B)}$. In our case, $P(A) = P(X_{j1} = 1)$ and $P(B) = P((X_{j2} = X_{j3} = \cdots = X_{jk} = 1)$. We know that $P(X_{j1} = 1) = (1-\frac{1}{n})^n$. The intersection of event $P(X_{j1} = 1)$ and $P((X_{j2} = X_{j3} = \cdots = X_{jk} = 1)$ is $P(X_{j1} = X_{j3} = \cdots = X_{jk} = 1)$ as it means that bins $1, 2, \ldots k$ are empty. Consider,

$$P(X_{j1} = 1|P((X_{j2} = X_{j3} = \cdots = X_{jk} = 1) = \tfrac{P(X_{j1}=X_{j3}=\cdots=X_{jk}=1)}{P((X_{j2}=X_{j3}=\cdots=X_{jk}=1)}$$
$$= \frac{(\frac{n-k}{n})^n}{(\frac{n-(k-1)}{n})^n} = \frac{(n-k)^n}{(n-k+1)^n} = \big(1-\frac{1}{n-k+1}\big)^n < \big(1-\frac{1}{n}\big)^n$$
$$\text{Therefore, } P(X_{j1} = 1|P((X_{j2} = X_{j3} = \cdots = X_{jk} = 1) \leq P(X_{j1} = 1)$$

So using the statement above, we can say that,

$$P(X_{j1} = 1|P(X_{j2} = X_{j3} = \cdots = X_{jk} = 1) = \frac{P(X_{j1} = X_{j3} = \cdots = X_{jk} = 1)}{P(X_{j2} = X_{j3} = \cdots = X_{jk} = 1)} \leq P(X_{j1} = 1) \leq 1/e$$
$$\frac{P(X_{j1} = X_{j3} = \cdots = X_{jk} = 1)}{P(X_{j2} = X_{j3} = \cdots = X_{jk} = 1)} \leq 1/e$$
$$P(X_{j1} = X_{j3} = \cdots = X_{jk} = 1) \leq P(X_{j2} = X_{j3} = \cdots = X_{jk} = 1)/e$$

The probability $P(X_{j2} = X_{j3} = \cdots = X_{jk} = 1)$ can be written as $\frac{1}{e^{k-1}}$ by breaking it down into conditional probabilities ans carrying it out $k-1$ times. Therefore,

$$P(X_{j1} = X_{j3} = \cdots = X_{jk} = 1) \leq \tfrac{1}{e^k}$$

The probability that 90% of bins are empty can be calculated by plugging in $k = 0.9n$ in the above formula. Therefore, the required probability is $e^{-09.n} = (e^{-0.9})^n \approx (0.4)^n < (0.9)^n$

2: Estimating the Mean and Median

(a) We can use the Hoeffding's bound to the required probability which states that for independent random variables in the range $[s_i, r_i]$, the probability,

$$P(|\bar{X} - E[X]| \geq t) \leq 2 \cdot exp\big(-\tfrac{2n^2t^2}{\sum_{i=1}^{n}(r_i - s_i)^2}\big)$$

where X is the sum of independent random variables $X_1, X_2, \ldots X_n$ and $\bar{X}$ is sampled mean over the n samples.

We can use this formula as we have j samples and each of the j samples can have values only from the interval $[-1, 1]$. For our case $\bar{X}$ is $\hat{\mu}$, the mean of the sample mean and $E[X]$ is just the actual mean μ. Also t si just replaced by ϵ. By plugging appropriate variables to solve our problem, we get,

$$P(|\hat{\mu} - \mu| \geq \epsilon) \leq 2 \cdot exp\big(-\tfrac{2j^2\epsilon^2}{\sum_{i=1}^{j}(1-(-1))^2}\big)$$

$$P(|\hat{\mu} - \mu| \geq \epsilon) \leq 2 \cdot exp\big(-\tfrac{2j^2\epsilon^2}{4j}\big) = 2 \cdot exp\big(-\tfrac{j\epsilon^2}{2}\big)$$

Now $P(|\hat{\mu} - \mu| \leq \epsilon) = 1 - P(|\hat{\mu} - \mu| \geq \epsilon)$. Therefore,

$$P(|\hat{\mu} - \mu| \leq \epsilon) \geq 1 - 2 \cdot exp\big(-\tfrac{j\epsilon^2}{2}\big)$$

We want the L.H.S of the above inequality to be greater than $1 - \delta$. Therefore we should have,

$$1 - 2 \cdot exp\big(-\tfrac{j\epsilon^2}{2}\big) \geq 1 - \delta$$

$$2 \cdot exp\big(-\tfrac{j\epsilon^2}{2}\big) \leq \delta$$

Solving this for j gives,

$$j \geq \frac{2}{\epsilon^2} \ln\big(\frac{2}{\delta}\big)$$

This R.H.S should be the minimum sampling size to obtain the required bounds.

(b) If the sampling is done wihtout replacement then we cannot use Hoeffding's inequality due to the fact that $X_1, X_2 \ldots X_j$ don't reamin independent anymore. We have obtained X_i by picking from the $a_1, a_2 \ldots a_n$. If there is no replacement then the value X_i will depend on X_{i-1}, i.e., the value of the current sample being picked will depend on the last number picked as the last number can't be picked now. This affects the probability of value that will be assigned to current sample.

(c) If the numbers $a_1, a_2 \ldots a_n$ take values from the interval $[-M, M]$, we can use the Hoeffding's inequality but with values for $r_i = M$ and $s_i = -M$. By plugging in these values we get,

$$j \geq \frac{2M^2}{\epsilon^2} \ln\left(\frac{2}{\delta}\right)$$

Therefore, we can say that if weaken the constraint by M, we need to atleast M^2 times more samples to satisfy the given constraints.

(d) The actual median of the n numbers, $a_1, a_2 \ldots a_n$ will be the value of $(n + 1/2)^{th}$ element. Let's call this value m. Therefore, we want the median of the sample j should be a value in the range $(m - \epsilon, m + \epsilon)$. Consider the three regions, R_L, R_M, R_R, where R_L is the region that has elements from $a_1, a_2 \ldots a_n$ with value less than $(m - \epsilon)$, R_M is the region that has elements with values in the range $(m - \epsilon, m + \epsilon)$, R_R is region that contains the remaining elememts. Now when we pick the elements for our j-sized sample, if we pick $j/2$ elements from either R_L or R_R, then the median of this sample will be a element from these regions respectively. We just want opposite of these cases i.e. that out j values we pick $j/2$ should come from the region R_M.

3: Quick-sort with Optimal Comparisons

(a) The pivot that we select in this problem is the median of the sample $M = 2m + 1$. Therefore, as far as the sample is concerned, the pivot will have m elements smaller than its value and m elements larger than its value. A sample of size $2m + 1$ can be picked from n elements in $\binom{n}{2m+1}$ ways. We want k^{th} smallest element to be pivot, therefore, m elements should be picked from $k - 1$ smallest elements in the array and exactly m should be picked from $n - k$ largest elements in the array. The respective ways of selecting these elements is $\binom{k-1}{m}$ and $\binom{n-k}{m}$. The k^{th} smallest element in the array can only be picked in one way. Therefore, a m-sized sample can be picked in $\binom{k-1}{m} \cdot 1 \cdot \binom{n-k}{m}$ ways such that k^{th} smallest element in the array is median of this sample.
Therefore, $p_k = \dfrac{\binom{k-1}{m} \cdot \binom{n-k}{m}}{\binom{n}{2m+1}}$

(b) When we first select a pivot, we will do $n - 1$ comparisons to find the correct position of the pivot in the sorted array. This way we will have some elements to the left of this pivot and some elements to the right of the pivot. Next we will again select two pivots for each of the two regions and recurse. The expected number of comparisons is equivalent to the expected running time and can be written as,

$$T(n) = T(n - 1) + \text{Expexted cost of recursion}$$

A pivot is selected with the probability p_i and incurs a further cost $T(i - 1)$ and $T(n - i)$. Also note that the way select a pivot restricts the value of i in the range $[m + 1, n - m]$ (i can only take integer values) because there should be atleast m elements smaller than i^{th} element and $n - m$ elements larger than the i^{th} element. Therefore, the expected cost can be written as,

$$\text{Expexted cost of recursion} = \textstyle\sum_{i=m+1}^{n-m} p_i \cdot (T(i - 1) + T(n - i))$$

Therefore, the expected number of comparisons can be written as,

$$T(n) = T(n - 1) + \textstyle\sum_{i=m+1}^{n-m} p_i \cdot (T(i - 1) + T(n - i))$$

(c) We're given the approximation,

$$p_i = \frac{(2m+1)!}{m!m!} \cdot \frac{1}{n} \cdot \left(\frac{i}{n}\right)^m \cdot \left(1 - \frac{i}{n}\right)^m$$

From part(b), we have,

$$T(n) = T(n-1) + \sum_{i=m+1}^{n-m} \frac{(2m+1)!}{m!m!} \cdot \frac{1}{n} \cdot \left(\frac{i}{n}\right)^m \cdot \left(1 - \frac{i}{n}\right)^m \cdot \Big[T(i-1) + T(n-i)\Big]$$

This summation can be converted to an integral using $\sum_{i=1}^{n} f(i) \leq \int_1^n f(x)dx$

$$T(n) \leq T(n-1) + \frac{(2m+1)!}{m!m!} \cdot \frac{1}{n^{2m+1}} \int_{m+1}^{n-m} (x)^m \cdot (n-x)^m \cdot [T(x-1) + T(n-x)]$$

Let's assume that $T(i) \leq C \cdot i \log i$. For $i = 1, T(i) = 0$ i.e. we don't need to make any comparison for a single element array. Now for $m = 1$, the above inequality can be written as,

$$T(n) \leq T(n-1) + 3! \cdot \frac{1}{n^3} \int_2^{n-1} (x) \cdot (n-x) \cdot [(x-1)\log(x-1) + (n-x)\log(n-x)]$$

Solving this integral using sagemath, gives the largest growing expression $n^4 \log(n-2)$ (coefficient $= 1$) which when divided by n^3 gives $n \log(n)$ (we need the asymptotic behavior so removing the -2 term). Now if we replace x with $x+1$, it's easy to see that the largest growing term will still be the order of $n \log n$ (as we can use subsitution and change the limits). Hence, from induction we can say that $T(n) \leq C \cdot n \log n$.

This constant will be dependent on the value of m. Solving this integral for $m = 5$ gives the largest growing term as $n \log(n-6)$ (coefficient $= 1$).

For both cases, we have the bound of type $\log n - a$. Inorder to convert $\log(n-2)$ and $\log(n-6)$ to $log(n)$, the coefficient by which we have to divide $\log(n)$ is larger for the case of $(n-6)$ than $(n-2)$. Therefore, we can say that $C_5 < C_1$. This is equivalent to showing that $\frac{\log(n-6)}{\log(n)} < \frac{\log(n-2)}{n}$ (strict inequality).

Also the value of C_5 is dictated by the ratio of $\frac{\log(n-6)}{\log n}$, this value will never be greater than and hence ≤ 1.6.

4: Randomized Min-Cut

(a) E' is the given min-cut which divides the graph into set of nodes S and T ($S \cup T = V$) and the number of edges between these two set is $|E'|$. Any two nodes whose edge do not belong to E' must be placed either in S or T (it's not possible to have one node in S and other in T because then their edge is in min-cut). Let's say that we collapse the edge (u, v) which does not belong to E'. When we collapse an edge we keep its end-points in either S or T. This means that the edge connecting them will not counted in $|E'|$ and hence the the value of $|E'|$ remains same.

(b) Consider a cut S and $V - S$ such that S only one node u and $V - S$ contains all other nodes. Then the size of cut is simply $degree(u)$. Since the minimum cut E' is the smallest cut, its size must be smaller than this cut. Therefore,

$$\text{size of min cut} \leq degree(u)$$

This inequality holds for all the vertices in the graph. If we add all these inequalities, we get $n\cdot$(size of the min cut) $= n|E'|$ and $\sum_{u \in V} degree(u)$ on the right hand side. This summation is equal to $2|E|$. Dividing both sides by n gives us,

$$|E'| \leq 2|E|/n$$

(c) When we collapse the first edge, we don't want it to be part of the min-cut. Since, the size of the min-cut can be atmost $2|E|/n$. The probability that the first edge picked is part of the min-cut is $\frac{(2|E|/n)}{|E|} = \frac{2}{n}$ (because there are $|E|$ edges to pick from). With this maximum probability, our min-cut value will change after first pick, which we don't want. Therefore, the probability that the first edge doesn't belong to min-cut is atleast $(1 - 2/n)$.

Therefore, we can find the probability of success as follows,

$$\begin{aligned} P[success] &= P[\text{first edge is not in min-cut}] \cdot P[\text{second edge is not in min cut}] \cdot \ldots \\ &= P[success] = (1 - \tfrac{2}{n}) \cdot (1 - \tfrac{2}{n-1}) \cdot \ldots \end{aligned}$$

This product can go upto the point when the min-cut is actually the last edge only. In that case our last pick would when only 3 nodes (and edges) are left. Therefore,

$$P[success] \leq (1 - \tfrac{2}{n}) \cdot (1 - \tfrac{2}{n-1}) \cdot \cdots \cdot (1 - \tfrac{2}{3}) \approx \tfrac{2}{n^2}$$

(d) Let's say that are m min cuts in the graph denoted by $C_1, C_2, \ldots, C_m$ and $x_1, x_2, \ldots, x_m$ be the event that the algorithm produces that output. The probability of each of these events is atleast $\frac{2}{n^2}$.

Different run of Krager's algorithm can produce any of these min-cuts out of all the possible cuts (including these and other cuts). Since a particular output of the algorithm doesn't depend on another output, we can write,

$$\sum_m P[x_i] \leq 1$$

$$\sum_m (\tfrac{2}{n^2}) \leq 1$$

$$m \cdot \tfrac{2}{n^2} \leq 1$$

$$m \leq \tfrac{n^2}{2}$$

Therefore, the min-cuts in a graph is atmost $\frac{n^2}{2}$.

5: Valiant-Vazirani Lemma

The base case is when $m = 1$. In that case the given formula evaluates to 1. This is true beacause there will be always such a unique element which is the single element itself. So the expression holds for the base case.

Let's assume that the formula is true for $m-1$ elements, i.e. for $m-1$ elements the argmin exists with a probability of atleast $(1-\frac{1}{N})^{m-2}$ and let's call the value at the index argmin, v. For the case when there are m elements, the $(a_m)^{th}$ element can either be greater than, equal to or smaller than v. If it equal to v, then this is a bad case and we no longer have a unique argmin. If the new element is smaller than v, then the index of the new element becomes the new argmin. If the new element is greater than v then the previous argmin still remains the argmin. Therefore, the only bad case is when the new element is equal to v. Probability that the new element is not equal to v is $(1-\frac{1}{N})$. Therefore, the probability that there exists a unique argmin at m^{th} step is product of the probability that there is unique argmin at $(m-1)^th$ step and the next element still maintains this uniqueness. Therefore,

$$P[\text{argmin exists at } m^{th} \text{ step given that it exist at } (m-1)^th \text{ step}] \leq (1-\tfrac{1}{N})^{m-2} \cdot (1-\tfrac{1}{N})$$

$$= (1-\tfrac{1}{N})^{m-1}$$