1: Balls and Bins

(a) for all of the first question, we will use the relation that the probability of any given bin being empty is given as $(1-\frac{1}{n})^n \leq \frac{1}{e}$ which tends to $\approx \frac{1}{e}$ when n tends to ∞. we need to find the probability that a given bin is empty for 1 ball throw which is given as $(1-\frac{1}{n})$, for m ball throws (m being $4nlogn$ here, wherein log is natural log ln) the probability that any given bin is empty is given as

$= (1-\frac{1}{n})^m$

$= (1-\frac{1}{n})^{4nlogn}$

$= ((1-\frac{1}{n})^n)^{4logn}$

$= (\frac{1}{e})^{logn^4}$

$= (e)^{-logn^4}$

$= \frac{1}{n^4}$ which is $< \frac{1}{n}$

(a)

a) for $m = \frac{1}{2}nlogn$ the probability that m ball throws resulting in a given bin being empty is given as

$= (1-\frac{1}{n})^{\frac{1}{2}nlogn}$

$= (1-\frac{1}{n})^{nlog\sqrt{n}}$

$= ((1-\frac{1}{n})^n)^{log\sqrt{n}}$

$= (\frac{1}{e})^{log\sqrt{n}}$

$= (e)^{-log\sqrt{n}}$

$= \frac{1}{\sqrt{n}}$, this signifies that the probability of any given bin being empty is more as compared to the first case as the empty probability is $> \frac{1}{n^4}$

b) for $m = 100nlogn$ the probability that m ball throws resulting in a given bin being empty is given as

$= (1-\frac{1}{n})^{100nlogn}$

$= (1-\frac{1}{n})^{nlogn^{100}}$

$= ((1-\frac{1}{n})^n)^{logn^{100}}$

$= (\frac{1}{e})^{logn^{100}}$

$= (e)^{-logn^{100}}$

$= \frac{1}{n^{100}}$, this signifies that the probability of any given bin being empty is very less as compared to the first case as the empty probability is $<<< \frac{1}{n^4}$

(c) Using Markov's inequality to bound the relation which is given as

$Pr(x \geq a) \leq \frac{E(x)}{a}$

for n bins being empty, the expected number of empty bins is given as $n(1-\frac{1}{n})^n \approx \frac{n}{e}$. if we relate the above two equations, we can consider x as the event of having n empty bins, while this expectation can be related as $E(x) = \frac{n}{e}$. we can consider a as a condition where we have more than 90% of bins being empty. putting all equation together -

$= Pr(x \geq 90\% of n)$

$= Pr(x \geq \frac{9}{10}n)$

$= Pr(x \geq \frac{9e}{10}\frac{n}{e})$

$= Pr(x \geq \frac{9e}{10}E(x))$ // applying Markov's equation from above

$\leq \frac{10}{9e}$

≈ 0.4

(d) from the given expression to prove, let X_j denote the random variable whose value is given as -

$$X_j = \begin{cases} 1 \text{ if bin } j \text{ is empty} \\ 0 \text{ otherwise} \end{cases}$$

let A denote the event that all the k bins from $j1, j2, j3, j4...jk$ are empty, whose probability can be written as -

P(A) $= P[X_{j1} = X_{j2} = X_{j3} = ... = X_{jk} = 1] = (\frac{n-k}{n})^n$

let B denote the event that the bins apart from $j1$ which are $j2, j3, j4...jk$ being empty, whose probability can be written as -

P(B) $= P[X_{j2} = X_{j3} = ... = X_{jk} = 1] = (\frac{n-(k-1)}{n})^n = (\frac{n-k+1}{n})^n$ // as k is being reduced by 1

We know that the probability of occurrence of A depending on probability outcome of B can be written as -

P(A/B) $= P[X_{j1} = 1 \mid X_{j2} = X_{j3} = ... = X_{jk} = 1]$

From conditional probability -

P(A/B) $= \frac{P(A \cap B)}{P(B)}$

$P[X_{j1} = 1 \mid X_{j2} = X_{j3} = ... = X_{jk} = 1] = \frac{P[X_{j1}=X_{j2}=X_{j3}=...=X_{jk}=1]}{P[X_{j2}=X_{j3}=...=X_{jk}=1]}$

$P[X_{j1} = 1 \mid X_{j2} = X_{j3} = ... = X_{jk} = 1] = \frac{(\frac{n-k}{n})^n}{(\frac{n-k+1}{n})^n} = (\frac{n-k}{n-k+1})^n = (1 - \frac{1}{n-k+1})^n \leq (1-\frac{1}{n})^n$

hence, $P[X_{j1} = 1 \mid X_{j2} = X_{j3} = ... = X_{jk} = 1] < P[X_{j1} = 1]$ which proves the first part of the question.

from the first equation in part a we know that $(1-\frac{1}{n})^n \leq \frac{1}{e}$ which tends to $\approx \frac{1}{e}$ when n tends to ∞

comparing the equation with what we proved now, we can write -

$\frac{P[X_{j1}=X_{j2}=X_{j3}=...=X_{jk}=1]}{P[X_{j2}=X_{j3}=...=X_{jk}=1]} \leq \frac{1}{e}$

$P[X_{j1} = X_{j2} = X_{j3} = ... = X_{jk} = 1] \leq \frac{1}{e}P[X_{j2} = X_{j3} = ... = X_{jk} = 1]$ —- (1)

based on the above (1) equation we can write similar equations for the terms following up after $j2$

$P[X_{j2} = X_{j3} = ... = X_{jk} = 1] \leq \frac{1}{e}P[X_{j3} = X_{j4} = ... = X_{jk} = 1]$ and so on...

after k terms are exhausted, we can rewrite the equation (1) as

$P[X_{j1} = X_{j2} = X_{j3} = ... = X_{jk} = 1] \leq \frac{1}{e}\frac{1}{e}....\frac{1}{e}$//k times

$P[X_{j1} = X_{j2} = X_{j3} = ... = X_{jk} = 1] \leq \frac{1}{e^k}$ —- (2)

we can consider equation (2) as the new bound for k bins being empty which is much tighter than the previous Markov's claim.

For 90% of bins to be empty among n bins, we have -

$k = 90\%\text{of } n = (0.9n)$

re-substituting k value in equation (2).

P[90% bins being empty] $\leq \frac{1}{e^{0.9n}} = (e^{-0.9})^n = (0.406)^n$

thus, P[90% bins being empty] $= (0.406)^n < (0.9)^n$

2: Estimating the mean and median

(a) based on the interval bound mentioned, we can use one of the clauses of Hoeffding's theorem which states that if $X_1, x_2...X_n$ are some random independent random variables bound by any interval $[r_i, s_i]$, then

$$\text{P}\Big(\mid \overline{X} - E[\overline{X}] \mid \geq t \Big) \leq 2.e^{-\frac{2n^2t^2}{\sum_{i=1}^{n}(r_i - s_i)^2}}$$

from the equation, for i = 1 to n, X_i is the random variable for each sample, $\overline{X}$ is the sampled mean over n samples, while $E[\overline{X}]$ represents the Expected mean over n samples. For our case, we can relate X as μ, ϵ as t, samples $n = j$, and interval range $[r_i, s_i] = [-1, 1]$.

substituting this in above equation -

$$\text{P}\Big(\mid \hat{\mu} - E[\hat{\mu}] \mid \geq \epsilon \Big) \leq 2.e^{-\frac{2j^2\epsilon^2}{\sum_{i=1}^{j}(1-(-1))^2}} = 2.e^{-\frac{j\epsilon^2}{2}}$$

subtracting 1 from each side which will reverse the polarity of the main expression

$$\text{1-P}\Big(\mid \hat{\mu} - E[\hat{\mu}] \mid \geq \epsilon \Big) \geq 1 - 2.e^{-\frac{j\epsilon^2}{2}}$$

$$\text{P}\Big(\mid \hat{\mu} - E[\hat{\mu}] \mid \leq \epsilon \Big) \geq 1 - 2.e^{-\frac{j\epsilon^2}{2}}$$

as per the requirement in the we need to have a upper bound at ϵ while it needs to be atleast $1 - \delta$ on the probability. Thus, the RHS of the expression needs to have a value greater than equal to $1 - \delta$

explicitly writing the above condition -

$1 - 2.e^{-\frac{j\epsilon^2}{2}} \geq 1 - \delta$

$\delta \geq 2.e^{-\frac{j\epsilon^2}{2}}$

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

$j \geq \frac{2}{\epsilon^2}ln(\frac{2}{\delta})$, here j is the sample index such that the sampled mean and the expected mean difference is bounded by a small error ϵ with a probability $1 - \delta$.

(b) Hoeffding bound range are defined specifically for sampling with replacement. The above proof does not hold good if we sample without replacement, as the probabilities of the upcoming samples will change in that case.

(c) we will replace the range $[r_i, s_i] = [-M, M]$ in the above derived equation for probability which gives us -

$\mathrm{P}\Big(\mid \hat{\mu} - E[\hat{\mu}] \mid \leq \epsilon \Big) \geq 1 - 2.e^{-\frac{j\epsilon^2}{2M^2}}$

solving the equation in a similar way as part(a) gives us by taking the range bound of δ -

$1 - 2.e^{-\frac{j\epsilon^2}{2M^2}} \geq 1 - \delta$

$\delta \geq 2.e^{-\frac{j\epsilon^2}{2M^2}}$

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

$j \geq \frac{2M^2}{\epsilon^2} ln\frac{2}{\delta}$

from above equation we can see that the sample index size is directly proportional to the square of the sample range, i.e., sample size j increases as a square proportion term if range M increases, and while it decreases in square proportion if the range M decreases.

(d) Let A be the given dataset of size n and n being odd, we know that there will be exactly half the number of $(n-1)$ elements lying on the right side of the median n, and exactly half the number of $(n-1)$ elements lying on the left side of the median.From the given allowed error correction, we essentially want the median to lie within a correction factor of $(\frac{1}{2} - \epsilon).n \mapsto (\frac{1}{2} + \epsilon).n$ based on the number of odd elements, let's divide the dataset into three groups, Left (L), Middle(M), and Right(R). matching the ϵ correction factor from above into these three partitions, we can write for every element x_i such that -

$M = \{x \in A : \text{if position of } x \leq (\frac{1}{2} - \epsilon).n$

$L = \{x \in A : (\frac{1}{2} - \epsilon).n < \text{if position of } x < (\frac{1}{2} + \epsilon).n$

$R = \{x \in A : \text{if position of } x \geq (\frac{1}{2} + \epsilon).n$

In case of k samples being picked, if more than $\frac{k}{2}$ samples are from L, then the median will lie in L, and so is the case for R.

let X_j be a variable such that -

$X_j = \binom{1 \text{ with probability } (\frac{1}{2} - \epsilon)}{0 \text{ otherwise}}$ for L

$X_j = \binom{1 \text{ with probability } (\frac{1}{2} + \epsilon)}{0 \text{ otherwise}}$ for R

for any given partition,

$X = \sum_{j=1} kX_j$ denoted total samples and $E[XL], E[XR]$ denote the expectations for L and R respectively.

$E[XL] = k(\frac{1}{2} - \epsilon)$

$E[XR] = k(\frac{1}{2} + \epsilon)$

we need atleast $\frac{k}{2}$ elements to be in the respective domain for median to lie in the same split. Thus, bounding the probability as $X \geq \frac{k}{2}$

the equation can be re-written to suit the Chernoff bound approximation as -

$P[X \geq \frac{k}{2}] \leq P[X \geq (1+c)E[X]]$

which computes as -$c \leq \frac{2\epsilon}{1-2\epsilon}$

for $\epsilon = \frac{1}{8}$, we can approximate the range of C between $0 \mapsto \frac{1}{3}$

Now, using Chernoff Bound to approximate the probability of manufacturing function and the correction factor from above with bound $\geq \frac{k}{2}$

we can approximate the probability as -

$P_L(X \geq \frac{k}{2}) \leq e^{\epsilon^2(\frac{1}{2}-\epsilon)\frac{k}{3}}$

$P_R(X \geq \frac{k}{2}) \leq e^{\epsilon^2(\frac{1}{2}+\epsilon)\frac{k}{3}}$

couldn't take it forward from here as to how to prove whether a median could be found.

3: quick sort with optimal comparisons

(a) For a given dataset A with total size of n, and sample size $M = 2m + 1$, we have M as the number of random sample elements that we pick from A and $m > 1$. The calculated sample median element, is the median of (2m + 1) elements. Thus, it should have exactly m elements on either side from the set of random samples. If this element is the k^{th} smallest element in entire dataset A, that means, the m elements to the left of the pivot are chosen from the $(k-1)$ elements that are smaller than the median since it is the k^{th} smallest. Similarly, the other m elements to the right of the median are chosen from the $(n-k)$ elements that are greater than the median since it is the k^{th} smallest.

Now, putting the terms of k, n, and m in a choose combination function for probability gives us this -

$= Pr[\text{choosing} k^{th} \text{smallest element from sample } M]$

$= \frac{(\text{choosing } k^{th} \text{smallest from the left of pivot k in sample } M).(\text{choosing } k^{th} \text{smallest from the right of pivot k in sample } M)}{(\text{choosing total sample size M from the given total dataset})}$

$= \frac{(\text{choose } k-1 \text{elements from left m samples}).(\text{choose } n-k \text{elements from right m samples})}{(\text{choose n total samples from the given total dataset size } 2m+1)}$

$P[K^{th} pivot] = \frac{\binom{k-1}{m}\binom{n-k}{m}}{\binom{n}{2m+1}}$

(b) For given $2m + 1$ samples, we will have atleast m elements on its left, and m elements to it's right. For selecting a pivot, we need to perform a total of $(n1)$ comparison operations to create the left(L) set and the right(R) set, where the left set corresponds to elements less than or equal to the pivot and the right set corresponds to the elements greater than the pivot.

Now, based on the pivot ranging from $(m + 1)$ to $(n - m)$, we can split the dataset in sizes of

(m) and $(n - m - 1)$ or $(m + 1)$ and $(n - m - 2)$ or $(m + 2)$ and $(n - m - 3)$ and so on.

The pivot cannot go less than $(m + 1)$ or greater than (nm), because we select the pivot as the median of the $2m + 1$ samples. For the chosen pivot being the k^{th} smallest element, the recursive cost will be recursions into the arrays of sizes $(k - 1)$ and $(n - k)$.

Let X_k denote a random variable defined as

$X_k = \binom{T(k-1)+T(n-k) \text{ with probability } P_k}{0 \text{ with probability } 1-P_k}$

where $T(k-1)$ is the recursive cost on $(k-1)$ elements and $T(n-k)$ is the cost on remaining $(n-k)$ elements. Thus the expected number of comparison, $T(n)$, is the sum of the cost due to comparisons required to break-up the array into two sets for a chosen pivot, plus the expected Recursive cost.

Thus, we have -

$T(n) = (n-1) + E(\sum_k X_k)$

$$T(n) = (n-1) + \sum_{k=m+1}^{n-m} P_i[T(k-1) + T(n-k)]$$

(c)

we need to solve the recursion derived above in part (b) using the following recursion -

$p_k \approx \frac{(2m+1)!}{m!\ m!} \cdot \frac{1}{n} \left(\frac{k}{n}\right)^m \left(1 - \frac{k}{n}\right)^m$

applying it to the recursion value above, we will have -

$$T(n) = (n-1) + \sum_{k=m+1}^{n-m} P_i[T(k-1) + T(n-k)]$$

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

rearranging the terms in denominator for n

$$T(n) = (n-1) + \sum_{k=m+1}^{n-m} \frac{(2m+1)!}{(m!)^2 n^{2m+1}} (k)^m (n-k)^m [T(k-1) + T(n-k)]$$

As the pivots split in the middle, we need to intuitively come up with a constant of the form $cnlogn$ for some constant c. thus, we can try and write our recursion in an integral form as

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

The basic wolfram crashed and asked for a pro-license on the above expression evaluation. I could only approximate the evaluations based on the constant type present here and hence couldn't use any computing solvers to get the answers.

atleast for m=1, we can have -

$$T(n) \leq (n-1) + \frac{(3)!}{(1!)^2 n^3} \int_{2}^{n-1} x(n-x)[(x-1)log(x-1) + (n-x)log(n-x)]$$

to compute the equations on C_5 substituting m=5

$$T(n) \leq (n-1) + \frac{(11)!}{(5!)^2 n^{11}} \int_{6}^{n-5} x^5 (n-x)^5 [(x-1)log(x-1) + (n-x)log(n-x)]$$

approximating the equation by removing all the constant terms and reevaluating only using the predominant n terms.

we get $T(n)_{m=5} \leq (n-1) + nlog(n) \leq nlog(n) = C_5' nlog(n)$

thus, by property of induction, we can bound the comparisons for the above equation as $\leq C_m nlogn$

4: Randomized MinCut

For a given graph, $G = (V, E)$, a cut can be defined as -

$C(T) = (u, v) \in E, s.t. u \in T, v \in V - T$

(a) Let's say we have E' as the min-cut for the given graph $G = (V, E)$. Let us assume that it divides the vertices into 2 sets, T and $V - T$. we will discard every edge that collapses itself into a supernode(as defined in question), which forms a cycle corresponding to itself. Since the collapsed edge (u, v), does not belong to the min-cut, either distinctly or as a pair, they should belong to either T or $V - T$. Now, the collapse of nodes in (u, v) will result in a new supernode N' which will either be part of T or $V - T$, forming a cycle and hence will be discarded. In the end, edge (u, v) also gets discarded and does not contribute to the count of parallel edges, hence it does not affects the final min-cut. Thus collapse of an edge that does not belong to a min-cut does not modify the min-cut in the new sub graph.

(b) let X_i be a random variable such that

$X_i = \begin{pmatrix} \text{degree of vertex index 'i' with probability } P_i = \frac{1}{n} \\ 0 \text{ otherwise with probability } 1 - P_i \end{pmatrix}$

and let X be the summation of all degrees given as $X = \sum_i X_i$

Let A_n be the average degree of a given node over a computed mean,E' be the min-cut ,and D_i be the degree of each vertex i. We know that sum of all vertices of a graph is equal to twice the number of edges.

$A_n = E(X) = \sum_{i \in V} P_i.D_i = \sum_{i \in V} \frac{1}{n}.D_i = \frac{2|E|}{n}$

If we divide the graph into two sets such that first set contains a single vertex u, while the second set contains the rest of $(n - 1)$ vertices and we can safely assume that the degree of this cut can be D_u

Since this is a valid cut, we can write $\forall u \in V$

size of cut $=\mid E' \mid \leq min(D_u)$, now plugging results from the previous expression where we proved for average degree for a node, which is nothing but the RHS in the above expression.

thus, $\mid E' \mid \leq \frac{2|E|}{n}$

(c) From the above proof, let us assume that the algorithm returns a min-cut of size k. Thus we will have a sub-graph G' with atleast K order degree on the nodes. The input graph G on the other hand should have atleast $\frac{nk}{2}$ edges. Now using the proof of part (b) equation to write the remaining edges in original graph G.

we can have - $k \leq 2.\frac{|E|}{n}$

$\mid E \mid \geq \frac{nk}{2}$

Now, for the above graph, the probability of selecting one of the k edges to help form a min-cut can be written as -

$= \frac{k}{\frac{nk}{2}} = \frac{2}{n}$

the probability P after first step, where we have an edge selected which is not part of E', thus

maintaining the min-cut intact can be written as (where, let C' be the event for maintaining the min-cut across iterations.) -

$P[C'] = 1 - \frac{2}{n}$

Let's consider the same analogy for i step, we will be remaining with $(n-i)$ vertices and hence the number of edges retained can be given as $\frac{(n-i)k}{2}$.

From above two relations we can write -

$P[C'] = 1 - \frac{k}{\frac{(n-i)k}{2}} = 1 - \frac{2}{n-i}$

Now, in recursive iterations, counting only upto $n-3$ edges as mentioned for overall probability of success, we need to compute the product of all probabilities of $i = 0 \mapsto (n-3)$ edges wherein every iteration has to select an edge such that the original min-cut is maintained intact, and let E' be such an event to compute overall probability from the entire graph.

thus we can write - $P[E'] = \prod_{i=0}^{n-3} P[E_i] = (1 - \frac{2}{n})(1 - \frac{2}{n-1}).....(1 - \frac{2}{n-(n-4)})(1 - \frac{2}{(n-(n-3))})$

thus, $P[E'] = \frac{n-2}{n}.\frac{n-3}{n-1}.....\frac{1}{4}.\frac{1}{3} = \frac{2}{n(n-1)} \approx \frac{2}{n^2}$

(d) Let us assume there are k - E' min-cuts of the graph $G(V, E)$, say $E'_1, E'_2...E'_K$. we know that every individual iteration on algorithm produces one min-cut as output and let x_i be such event for i^{th} iteration of the algorithm. Also we showed in section (c) with Karger's algorithm, that the probability of single iteration output from the algorithm is $\approx \frac{2}{n^2}$ for all min-cuts, thus -

$P[E_i] \geq \frac{2}{n^2}$

Now, as the min-cut outputs are independent, we can write -

$\sum_{i}^{k} P[x_i] \leq 1$

now putting together all equation from above -

$\sum_{i}^{k} \frac{2}{n^2} \leq 1$

$K.\frac{2}{n^2} \leq 1$

$K \leq \frac{n^2}{2}$

5: Valiant Vazirani Lemma

(a) As hinted, we shall prove this by induction -

Let's build a base case to start off with. For a value of m=1, we have only 1 element chosen independently at random, thus it will always be the unique argmin since there are no other elements left in the chosen set.

for m=1, $(1 - \frac{1}{N})^{m-1} = 1$

now for $m \geq 1$, the probability that there exists a unique minimum, say U_m at m^{th} stage will always be $\geq (1 - \frac{1}{N})^{m-1}$

Now, to continue finding a new unique element, we will choose among the remaining $n-1$ elements a new m value as a'_m. If in case, the new element is $<$ current U_m, then we will have a

new U_m, else we will retain the previous value of U_m.

Now, the probability of $a'_m \neq U_m$ can be written as -

$P[a'_m \neq U_m] = \frac{N-1}{N} = (1 - \frac{1}{N})$

So, the probability that there exists a unique minimum U_m at every m^{th} stage is given as

$= P[\text{first } U_m]$ x $P[a'_m \neq U_m]$

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

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

Thus, by principle of induction, the condition holds true for all values.