CS6150 - Homework/Assignment-5

Arnab Das(u1014840)

November 20, 2016

1: Balls and Bins ...

We use the fact that $\lim_{n->\infty}(1-\frac{1}{n})^n \leq \frac{1}{e}$.

(a) Given m $= 4n\log n$ balls and n bins. We need to find the probability that atleast one bucket is empty.

Probability that bucket 'i' does not gets a ball in a single throw is $(1-\frac{1}{n}$. For m balls, then probability that the Bucket i is empty $= (1-\frac{1}{n})^m = (1-\frac{1}{n})^{4n\log n} = ((1-\frac{1}{n})^n)^{\log n^4} \leq \exp(-\log n^4) = \frac{1}{n^4} < \frac{1}{n}$.

(b)

(a): $m = \frac{1}{2}n\log n = n.\log(\sqrt[2]{n})$. Therefore, we get:

$(1-\frac{1}{n})^{n.\log(\sqrt[2]{n})} \leq \exp(-\log(\sqrt[2]{n})) = \frac{1}{\sqrt[2]{n}}$

(b): $m = 100n.\log n = n.\log(n^{100})$. Therefore, we get:

$(1-\frac{1}{n})^{n.\log(n^{100})} \leq \exp(-\log(n^{100})) = \frac{1}{n^{100}}$.

(c) Setting now, m=n. The probability that bin 'j' is empty is $(1-\frac{1}{n})^n \leq \frac{1}{e}$. Then the expected number of empty bins is $\frac{n}{e}$. If R denotes the event that a bin is empty, then we have got $E[R] = \frac{n}{e}$. To find the probability of 90% bins being empty, we can write it as follows:

$$P_r[R \geq 90\% of n] = P_r[R \geq \frac{9}{10}n] = P_r[R \geq \frac{9e}{10} \times \frac{n}{e}] = P_r[R \geq \frac{9e}{10}E[R]]$$

Now we can apply Markov's inequality to get:

$$P_r[R \geq \frac{9e}{10}E[R]] \leq \frac{10}{9e}$$

Thus, the probability that atleast 90% bins are emty is upper bounded by $\frac{10}{9e}$.

(d)

To Prove: $P_r[X_{j1} = 1 \Big| X_{j2} = X_{j3} = \cdots = X_{jk} = 1] \leq P_r[X_{j1} = 1]$.

Proof: Let X_j denote the random variable that is 1 if bin j is empty and 0 otherwise.

Let B denote the event that bins $j_2, j_3, \ldots, j_k$ are empty, then P(B) $= P[X_{j2} = X_{j3} = \cdots = X_{jk} = 1] = \left(\frac{n-k+1}{n}\right)^n$

Let A denote the event that bin j_1 is empty, then P(A) $= P[X_{j1} = 1] = \left(\frac{n-1}{n}\right)^n$ Then, $P(A \cap B)$ denote the event that bins $j_1, j_2, \ldots, j_k$ are empty, then P(A) $= P[X_{j1} = X_{j2} = \cdots = X_{jk} = 1] = \left(\frac{n-k}{n}\right)^n$

And,$P(A/B) = P[X_{j1} = 1 \Big| X_{j2} = X_{j3} = \cdots = X_{jk} = 1]$

By Conditional probability, we know:

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

$$P[X_{j1} = 1 \Big| X_{j2} = X_{j3} = \cdots = X_{jk} = 1] = \frac{P[X_{j1} = X_{j2} = \cdots = X_{jk} = 1]}{P[X_{j2} = X_{j3} = \cdots = X_{jk} = 1]}$$

$$P[X_{j1}=1\Big|X_{j2}=X_{j3}=\cdots=X_{jk}=1]=\frac{\left(\frac{n-k}{n}\right)^n}{\left(\frac{n-k+1}{n}\right)^n}=\left(\frac{n-k}{n-k+1}\right)^n=\left(1-\frac{1}{n-k+1}\right)^n\leq\left(1-\frac{1}{n}\right)^n$$

$$P[X_{j1}=1\Big|X_{j2}=X_{j3}=\cdots=X_{jk}=1]<P[X_{j1}=1]$$

(Proved).

Now, $\left(1-\frac{1}{n}\right)^n\leq\frac{1}{e}$ for $n->\infty$. From the above proof, we can write:

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

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

We can further break the brobability term on the rhs of equation(1) as:

$$\frac{P[X_{j2}=X_{j3}=\cdots=X_{jk}=1]}{P[X_{j3}=X_{j4}=\cdots=X_{jk}=1}]\leq\frac{1}{e}$$

$$\frac{P[X_{j3}=X_{j4}=\cdots=X_{jk}=1]}{P[X_{j4}=X_{j4}=\cdots=X_{jk}=1}]\leq\frac{1}{e}$$

and so on, to pull out the $\frac{1}{e}$ terms until we reach $P[X_{jk}=1]$. Clubbing all these together in equation(1), we get:

$$P[X_{j1}=X_{j2}=\cdots=X_{jk}=1]\leq\frac{1}{e^k} \tag{2}$$

(Intermediate Proof.) Equation(2) bounds the probability of k bins being empty as $\frac{1}{e^k}$. For n bins, to find the probability that 90% of the bins are empty, means $k=90\%ofn=(0.9)n$
P[90% bins are empty] $\leq e^{-(0.9)n}=(e^{-0.9})^n=0.4065^n<(0.9)^n$

2: Estimating the Mean and the Median ...

(a) Let $a_1,a_2,\ldots,a_n$ be n real numbers in $[-1,1]$. Given, μ is the expected mean, and $\hat{\mu}$ is the sampled mean. We need to find the sample index 'j', such that the sampled mean and the expected mean difference is bounded by a small error ϵ with a probability $1-\delta$.
The theorem 2 of Hoeffding states that, if $X_1,X_2,\ldots,X_n$ be independent random variables strictly bounded by the intervals $[r_i,s_i]$, then

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

Where $X=\sum_{i=1}^n X_i$, and X_i are the random variable that has the result of each random sample, $\bar{X}$ is the sampled mean over n samples and $E[\bar{X}]$ is the theoretical or expected mean of the distribution.
Relating this to our problem, over 'j' samples, we have samples mean as $\hat{\mu}$ and the theoretical mean as μ, the error t as ϵ with the strict bound interval for each sample as $[-1,1]$ such that $\forall i, r_i=-1 and s_i=1$. Thus,fitting this data into the Hoeffding inequality, we get:

$$P\bigg(|\hat{\mu}-\mu|\geq\epsilon\bigg)\leq 2\times\exp\bigg(-\frac{2j^2\epsilon^2}{\sum_{i=1}^j(1-(-1))^2}\bigg)=2\times\exp\bigg(-\frac{j\epsilon^2}{2}\bigg)$$

$$1 - P\bigg(|\hat{\mu} - \mu| \geq \epsilon\bigg) \geq 1 - 2 \times \exp\bigg(-\frac{j\epsilon^2}{2}\bigg)$$

$$P\bigg(|\hat{\mu} - \mu| \leq \epsilon\bigg) \geq 1 - 2 \times \exp\bigg(-\frac{j\epsilon^2}{2}\bigg) \tag{3}$$

We want the probability error bound of the difference of sampled and theoretical mean to be upper bounded by ϵ with atleast probability of $1 - \delta$. Hence, in equation(3), if we bound the rhs by $1 - \delta$, the lhs is also lower bounded by $(1 - delta)$, as per our requirement. Hence,

$$1 - 2 \times \exp\bigg(-\frac{j\epsilon^2}{2}\bigg) \geq 1 - \delta$$

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

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

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

Equation(4) gives a minimal sampling size required to get the approximate $(1 - \delta)$ guarantee of an error bound of ϵ in the difference of the theoretical and the sampled mean.

(b) If we sample without replacement, then the above proof does not holds anymore, since Hoeffding bounds are defined specifically for sampling with replacement. Sampling without replacement, changes the probabilities of the future samples which causes the Hoeffding bounds to break down.

(c) If we weaken the constraint on the a_i to $a_i \in [-M, M]$, then the interval limits for each sample becomes, $r_i = -M$, and $s_i = M$. Then equation(3) becomes:

$$P\bigg(|\hat{\mu} - \mu| \leq \epsilon\bigg) \geq 1 - 2 \times \exp\bigg(-\frac{j\epsilon^2}{2M^2}\bigg) \tag{5}$$

Rearranging the terms of equation(5) in a similar way as done for equation(3), we get:

$$j \geq \frac{2M^2}{\epsilon^2} ln(\frac{2}{\delta}) \tag{6}$$

Thus, if the interval size increases or decreses by the parameter M, the minimum sample size to attain the required guarantees also increases or decreases respectively, as the square of the parameter M.

(d) For the given dataset, $S = [a_1, a_2, \ldots, a_n]$, we want to determine the ϵ estimated medianby computing the sampled median. We define rank of the elements in the data set as:

$$rank(x) = |\{a \in S : a \leq x\}|$$

For a given odd dataset of size $n = 2m + 1, m \geq 0$, the actual median will have exactly m elements on either side. We can consider that we want the ϵ estimated median to be within $(\frac{1}{2} - \epsilon)n$ to $(\frac{1}{2} + \epsilon)$. So, we partition S into 3 groups as below:

$$S_L = \{x \in S : rank(x) \leq (\frac{1}{2} - \epsilon)n\}$$

$$S_M = \{x \in S : (\frac{1}{2} - \epsilon)n < rank(x) < (\frac{1}{2} + \epsilon)n\}$$

$$S_U = \{x \in S : rank(x) \geq (\frac{1}{2} - \epsilon)n\}$$

Observe that for t samples, if more than $\frac{t}{2}$ are from S_L, then the median will be in S_L. Similarly, if more than $\frac{t}{2}$ are from S_U, then the median will be in S_U. This implies that if less than $\frac{t}{2}$ samples are from S_L or S_U, then the sampled median will be an ϵ approximate median. Lets consider case by case for S_L and S_U.
In case of S_L: Let us define random variable, X_i, for each sample, such that, it is 1 if selected from S_L and 0 otherwise: Hence:

$X_i = 1$ with probability $p_i = (\frac{1}{2} - \epsilon)$

$X_i = 0$ otherwise

Then , $X = \sum_i X_i$, denotes the number of samples from S_L.

$$E(X) = \sum_i E(X_i) = (\frac{1}{2} - \epsilon)t$$

We would like to bound the probability of $X \geq \frac{t}{2}$, such that

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

which allows us to use the Cheroff bound . Here, we still need an estimate for the conatnt c. Note that, the above relation requires:

$$\frac{t}{2} \geq (1+c)E(X) = (1+c)(\frac{1}{2} - \epsilon)t$$

$$c \leq \frac{2\epsilon}{1-2\epsilon}$$

Notice that if ϵ increases then the upper bound of c increases, and if ϵ decreases then the upper bound of c decreases. For $\epsilon < \frac{1}{8}$, c has the range $0 \leq c \leq \frac{1}{3} < 1$. So, in the range for $\epsilon < \frac{1}{8}$, we can approximate c with ϵ.
So, using Chernoff bound we can write:

$$P_{S_L}(X \geq \frac{t}{2}) \leq P(X \geq (1+\epsilon)E(X)) \leq e^{-\epsilon^2(\frac{1}{2}-\epsilon)\frac{t}{3}}$$

In case of S_U Similar analysis for S_U, results in

$$P_{S_U}(X \geq \frac{t}{2}) \leq P(X \geq (1+\epsilon)E(X)) \leq e^{-\epsilon^2(\frac{1}{2}-\epsilon)\frac{t}{3}}$$

If we want the sampled median to be the ϵ estimated median with a high probability of $(1-\delta)$ for very small δ, that means the probability that median is in S_L or S_U should be bounded by δ. Hence,

$$P_{S_L}(X \geq \frac{t}{2}) + P_{S_U}(X \geq \frac{t}{2}) \leq 2e^{-\epsilon^2(\frac{1}{2}-\epsilon)\frac{t}{3}} \leq \delta$$

Rearranging the inequality, we get:

$$t \geq \frac{3}{\epsilon^2(\frac{1}{2} - \epsilon)} ln(2\delta^{-1})$$

For, a good guarantee requiring a small δ, say 0.01, will require $t \geq \frac{18}{\epsilon^2(\frac{1}{2}-\epsilon)}$ Notice , that t becomes inversly proportional to square of ϵ. Thus, for very small ϵ's $< 1/8$, the requires sample size for a good $(1-\delta)$ guarantee starts to increase. Since sampling is $o(n)$, then suppose sampling size, $t = \log n$. Then:

$$n \geq e^{\frac{3}{\epsilon^2(\frac{1}{2}-\epsilon)} ln(2\delta^{-1})}$$

which for very small ϵ and δ is an exponentially huge number, infeasible for normal datasets. Hence, it is not possible to estimate the median in this case for $t = \log n$.

Similarly, we can analyse for other $o(n)$, To generalize, we can express a function that is $o(n)$ as $n^{1-\gamma}$, where, $0 < \gamma \leq 1$. Then , we get the n to be atleast:

$$n \geq (\frac{3}{\epsilon^2(\frac{1}{2}-\epsilon)} ln(2\delta^{-1}))^{\frac{1}{1-\gamma}}$$

$$n \geq (\frac{3}{\epsilon^2(\frac{1}{2}-\epsilon)} ln(2\delta^{-1}))^{\alpha}$$

where $\alpha > 1$. Since, the terms inside the bracket in the rhs is $>> 0$, for small ϵ and small δ, and $\alpha > 1$, hence the entire rhs is a strictly increasing function, increasing as the inverse square times α of the small ϵ. Hence, for $\epsilon < 1/8$, we can have ϵ and δ that makes the minimal bound on the dataset size, for a sampling order in $o(n)$ to become extremely large, rendering it not possible. (Proved).

3: Quick-Sort with optimal Comparisons ...

(a) Given an input array, $A[0, \ldots, n-1]$, and we consider a variant of the quick-sort, where instead of picking a uniformly random pivot, we sample M random elements of A(without replacement), and pick the median of these entries without replacement, as the pivot. For $M = 2m + 1$, random elements that we pick from A and $m > 1$, the element that is chosen as the pivot, is the median of $(2m + 1)$, hence it has exactly m elements on either side from the set of random samples. If this pivot is the k'th smallest element in entire A, that means, the m elements to the left of the pivot is chosen from the (k-1) elemets that are smaller than the pivot since it is the k'th smallest. Similarly, the other m elements to the right of the pivot is chosen from the (n-k) elements that are greater than the pivot since it is the k'th smallest. The number of possible combinations of such a choice, for the k'th smallest being selected as the pivot is $\binom{k-1}{m} \times \binom{n-k}{m}$. The total number of choices of $2m + 1$ elements from the N size input array A, is $\binom{n}{2m+1}$. Hence, the probability, p_k, that the $k'th$ smallest element in A is chosen as the pivot will be :

$$p_k = \frac{\binom{k-1}{m}\binom{n-k}{m}}{\binom{n}{2m+1}} \tag{7}$$

(b) On selecting a pivot, we need to perform $(n-1)$ comparison operations to create the left(L) set and the right(R) set, where the left set corresponds to elements less than equal to the pivot and the right set corresponds to the elements greater than the pivot. Now, based on the pivot we may split the array 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 for the pivot ranging from $(m+1)$ to $(n-m)$. The pivot cannot go less than $(m+1)$ or greater than $(n-m)$, because we select the pivot as the median of the $2m+1$ samples, which means it will have atleast m elements on its

left and m elements on its right. For the chosen pivot being the i'th smallest element, the recursive cost will be recursions into the arrays of sizes (i-1) and (n-i). Let X_i denote a random variable defined as

$X_i = T(i-1) + T(n-i)$, with probability p_i

$X_i = 0$, with probability $1 - p_i$

where T(i-1) is the recursive cost on the array of (i-1) elements and T(n-i) is the recursive cost on the array of (n-i) elements. Thus the expected number of comparision,$T(n)$, is the sum of the cost due to comparisons required to break-up the array into L and R set for a chosen pivot, plus the Expected Recursive cost.

$$T(n) = (n-1) + ExpectedRecursiveCost, E(\sum_i X_i)$$

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

(c) Given the following approximation of p_i to solve the recursion:

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

Thus, we have from the previous recursion:

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

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

Now we can solve this by the "guess and prove inductively" method. Intuitively most pivots should split their array roughly in the middle, which suggests a guess of the form $cn\log n$ for some constant c. We should use this guess, we evaluate the resulting summation. We are guessing that $T(i) \leq ci\log i$ in the limits of the summation. For T(1), our guess works, since T(1) is 0. Then, we can write our recursion as some integral of the form:

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

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

We use "sageMath" software to solve this integral. Basic Wolfram crashed and asked for pro license. Thus, for **m=1**, the formation looks like:

$$T(n) \leq (n-1) + \frac{(3)!}{(1!)^2 n^3}\int_2^{n-1} x(n-x)[(x-1)\log(x-1) + (n-x)\log(n-x)]$$

The solution from SageMath looks like:

```
n = var('n')
f = x*(n-x)*((x-1)*log(x-1,2) + (n-x)*log(n-x,2))
assume(n-2>0)
f.integrate(x,2,n-1)
```

```
-1/72*(7*n^4 - 22*n^3 - 18*n^2 - 12*(n^4 - n^3 - 15*n^2 + 40*n - 28)*log(n - 2) + 78*n - 45)/log(2)
```

Figure 1: Sage solution for m=1

picking the *nlogn* term because that is the highest order term in the solution, and hence drives the number of comparisons.

$$T_{m=1}(n) \leq (n-1)+\frac{6}{n^3}(\frac{-1}{72}(-12n^4\log(n-2))) = (n-1)+n\log(n-2) \leq n\log(n-2) \leq n\log(n-2) = C_1'n\log(n-2)$$

Similarly, for **m=5**, the formation looks like:

$$T(n) \leq (n-1) + \frac{(11)!}{(5!)^2 n^{11}} \int_6^{n-5} x^5(n-x)^5[(x-1)\log(x-1) + (n-x)\log(n-x)]$$

The solution from SageMath looks like:

```
n = var('n')
f = x^5*(n-x)^5*((x-1)*log(x-1,2) + (n-x)*log(n-x,2))
assume(n-7>0)
f.integrate(x,6,n-5)
```

```
-1/76839840*(18107*n^12 - 159017*n^11 - 363132*n^10 - 1522290*n^9 - 7023555*n^8 - 34220340*n^7 - 172867464*n^6 + 396*n^5*
(8416600500*log(5) + 279124883) - 22275*n^4*(3773677600*log(5) + 73982609) + 385*n^3*(2248924293000*log(5) + 26951446157)
 - 3388*n^2*(1338899310000*log(5) + 9700044569) + 1386*n*(8696832165000*log(5) + 35833947371) - 27720*(n^12 - n^11 - 2877
3822*n^6 + 857971752*n^5 - 10914765345*n^4 + 75463373370*n^3 - 298039453866*n^2 + 635912430048*n - 571537891512)*log(n -
6) - 12967647348348000*log(5) - 25640761527381)/log(2)
```

Figure 2: Sage solution for m=5

Again, picking the *nlogn* term because that is the highest order term in the solution, and hence drives the number of comparisons.

$$T_{m=5}(n) \leq (n-1)+\frac{2772}{n^{11}}(\frac{-1}{76839840}(-27720n^{12}\log(n-6))) = (n-1)+n\log(n-6) \leq n\log(n-6) = C_5'n\log(n-6)$$

Thus, by principle of induction, the expected number of comparisons is bounded by $\leq C_m n\log n$.(Proved) Note that $\log(n) and \log(n-c)$ where c is a constant differ slightly in order and hence can be considered in order of logn.

Now, as we see,

$$T_{m=1}(n) \leq C_1'n\log(n-2) = C_1 n\log(n) - c_1$$
$$T_{m=5}(n) \leq C_5'n\log(n-6) = C_5 n\log(n) - c_5$$

where small c_1 and small c_5 are cosntants that we extract out to bring the terms to $\log n$ Now, $C_1' = C_5' = 1$, from our derivation. Since, C_1 comes by changing $\log(n-2)$ to $\log(n)$, the change(say $\alpha \geq 0$) is much **less**, than for C_5 which comes by changing(say $\beta \geq 0$) $\log(n-6)$ to $\log(n)$. Hence,

$$C_1' = C_1 + \alpha = C_5 + \beta = C_5'$$

since, $\alpha < \beta$, hence

$$C_1 > C_5$$

(Proved).

Also,

$$C_5 + \beta = C_5' = 1$$

for $\beta \geq 0$, hence:

$$C_5 \leq 1$$

Thus, $C_5 \leq 1 < 1.6$ (Proved).

4: Randomized Min-Cut . . .

Let the graph be $G = (V, E)$, where V is the set of vertices and E is the set of edges such that $n = |V|, m = |E|$. A Cut is defined as:

$$\delta(S) = (u, v) \in E : u \in S; v \in V - S$$

A Min-cut is $min\delta(S)$

(a) Prove: Prove that if we collapse an edge that is not in the min-cut E', then the size of the min-cut in the new graph is equal to G.

Proof: Suppose we are considering E' the min-cut for the given graph and the min-cut size is K, such that it divides the vertices into 2 sets S and $V - S$. Every edge, that collapses into a supernode, will form a self-loop on that supernode corresponding to itself, which we discard. Since, the collapsed edge, (u, v), does not belongs to a min-cut, that means the vertices u, and v, both together belongs to either S or $V - S$. Then, the collapse of (u, v) will result in a supernode, say w, which will be in the same set to which u and v belongs to, that is , either S or $V - S$, and this edge will form a self-loop on w, which will be discarded. Since, ultimately this edge (u, v), gets discarded and does not contributes to the count of parallel edges, hence it does not affects the final min-cut. Since, we made no assumptions for the E' min-cut, hence w.l.o.g this result holds true for every min-cut. Thus collapse of an edge that does not belongs to a min-cut does not changes the min-cut in the new graph. **Proved.**

(b) Prove: If E' is one of the min-cuts in the graph, prove that $|E'| \leq 2|E|/n$.

Proof: Let X_i denote a random variable, such that:

X_i = degree of vertex 'i' with probability $p_i = \frac{1}{n}$ $X_i = 0$ with probability $1 - p_i$ Let, $X = \sum_i X_i$, then the average degree of a node is the theoretical mean of the distribution given as:

$$Avgdegreepernode = E(X) = \sum_{i \in V} p_i degree(i) = \sum_{i \in V} \frac{1}{n} degree(i)$$

Now, we know that the sum of the degrees of all the vertices of a graph is equal to twice the number of edges in the graph. Then, we can write:

$$Avgdegreepernode = \frac{1}{n} \sum_{i \in V} degree(i) = \frac{1}{n} 2|E| \tag{8}$$

Now, consider a partition of V into two sets, one containing a single vertex, u, and the other containing the remaining (n-1) vertex. The size of this cut will be equal to the $degree(u)$. Since, this is valid cut, hence the mincut of this graph should be less than or equal to the $degree(u)$. This partition can be done for any vertex, and hence, the min-cut size must be less than equal to the $degree(u)$, $\forall u \in V$, Thus:

$$sizeOfMinCut \leq \min_{\forall u \in V} degree(u)$$

Since, the rhs is a minima overall all the vertices, hence it should be less than equal to the average degree of a node. Hence plugging in results from equation(8), we can write

$$sizeOfMinCut \leq \min_{\forall u \in V} degree(u) \leq \frac{2|E|}{n} \tag{9}$$

For the given min-cut, E', hence we can finally write:

$$|E'| \leq \frac{2|E|}{n} \tag{10}$$

(Proved).

(c) Suppose the edges we contact in the algorithm are $[e_1, e_2, \ldots, e_{n-2}]$. The algorithm successfully returns a min-cut, E', if none of these edges are in E'. Suppose the size of the min-cut is K. Then the minimum degree of the nodes in G is atleast K according to equation(9), else we can have a min-cut size less than K. Thus, the original input graph has atleast $\frac{nK}{2}$ edges since,

$$Min - Degree = K \leq AvgDegree = \frac{2|E|}{n}$$

$$|E| \geq \frac{nK}{2}$$

In part(a), we showed that every edge that is not part of a min-cut, if collapsed, then the size of the min-cut in the new graph, G' remains the same. This implies that there is a one to one correspondence between the cuts in G and the cuts in G'. Thus, in G', the minimum degree of value K is retained. Then for the first contraction, there are n vertices, and hence number of edges is $\frac{nK}{2}$. Probability of selecting one of the K-edges(which form the min-cut) is $\geq \frac{K}{\frac{nK}{2}} = \frac{2}{n}$.
Then, the probability that min-cut is maintained after first step is :

$$P_r(e_1 \notin E') = 1 - \frac{2}{n}$$

(Proved).

Consider the situation after j contractions. The number of remaining vertices will be $(n - j)$. Using equation(9), the number of remaining edges is atleast $\frac{(n-j)K}{2}$. Then the probability that even in the j'th step, min-cut if (j-1) steo is maintained :

$$P_r(e_j \notin E') = 1 - \frac{k}{(n-j)K/2} = 1 - \frac{2}{n-j}$$

Then the probability that Minimal cut of 1st step is maintained in the j'th step =

$$P_r(e_1 \notin E').P_r(e_2 \notin E').\ldots.P_r(e_j \notin E')$$

Thus, the probability that the final graph=E' will be:

$$P_r(FinalGraph = E') = \Pi_{j=0}^{n-3} P_r(e_j \notin E') \geq (1 - \frac{2}{n})(1 - \frac{2}{n-1}) \ldots (1 - \frac{2}{n-j}) \ldots (1 - \frac{2}{n-(n-3)})$$

$$P_r(FinalGraph = E') = \frac{n-2}{n}\frac{n-3}{n-1}\frac{n-4}{n-2} \cdots \frac{1}{3} = \frac{2}{n(n-1)} \approx \frac{2}{n^2}$$

(Proved).

(d) Suppose there are K min-cuts of the graph, say $C_1, C_2, \ldots, C_K$. Ley y_i be the event that C_i is output by the algorithm. Since, the min-cuts are output separately by the algorithm over different runs, hence these are disjoint events. For disjoint events, y_i, we know,

$$\sum_{i=1}^{K} P_r[y_i] \leq 1$$

In the last part, we showed that the probability that one run of Karger's algorithm outputs a min-cut is $\approx \frac{2}{n^2}$. This is equally likely for all the possible min-cuts, that is, for every min-cut, C_i, $P_r[C_i] \geq \frac{2}{n^2}$. Hence,

$$\sum_{i=1}^{K} P_r[y_i] \leq 1$$

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

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

$$K \leq \frac{n^2}{2} \tag{11}$$

Equation(11) shows that the number of min-cuts in a graph is $\leq \frac{n^2}{2}$.

(Proved).

5: Valiant-Vazirani Lemma

Let $a_1, a_2, \ldots, a_m$ be random integers chosen independently from the interval $[1, N]$. **Prove:** Probability that the argmin of the a_i(i.e., the index j such that a_j) is the minimum) is unique with probability atleast $(1 - \frac{1}{N})^{m-1}$.

Proof:

Statement **S1**: Probability that the argmin of the a_i is unique with probability atleast $(1 - \frac{1}{N})^{m-1}$. **BAse-case:** For m=1: if only 1 element is chosen independently at random, then it is the unique argmin always since there is no other remaining element in the chosen set. Hence, it is the unique argmin with probability 1. Solving $(1 - \frac{1}{N})^{m-1}$ for $m = 1$ is 1. Thus S1 holds true in the base case.

Assume: Assuming S1 is true for m. Then at the m'th stage there exists a unique minimum, say k, with probability $\geq (1 - \frac{1}{N})^{m-1}$.

The next randomly chosen integer is a_{m+1}. If $a_{m+1} \neq k$, then it can be either less than k, in which case it becomes the new unique minimum or it can be greater than k(may be unique or repeated), in which case our unique minimum from m'th stage, that is k, is retained. This choice of a_{m+1} can be done out of the other $(N-1)$ elements except k. Thus the probability that $a_{m+1} \neq k$ will be:

$$P_r(a_{m+1} \neq k)\frac{N-1}{N} = 1 - \frac{1}{N}$$

Hence , the probability that at the $(m+1)'th$ stage there exists a unqie minimum is
= (Probability that at the m'th stage there exists a unique minimum) $\times$ $P_r(a_{m+1} \neq k)$

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

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

Thus, S1 holds true for $(m+1)$ stage as well. Hence, by principle of induction, S1 is true for all integers in the interval $[1, N]$. (Proved).