1st Assignment

Ευστράτιος Ρέππας (el20002)

Exercise 1

In this exercise, we are asked to propose a probabilistic algorithm for the r-way Min-Cut problem. Since in the 2-way Min-Cut problem we tried contracting edges until 2 vertices remain, the idea is to generalize this approach by contracting till there are r remaining nodes, since they will represent the r components.

Then, we must calculate the probability of a randomly contracted edge belonging in C, where C is the set containing the edges that form the Cut. If this happens, we will end up with a suboptimal cut, so to calculate the probability of success, we will need to not choose such an edge for all $n - r$ iterations of the algorithm.

We start by assuming a graph $G(V, E)$and a set of vertices $V_s \subseteq V$, with cardinality $|A| = r - 1$. By removing all the edges that have an endpoint in V_s, we end up with an r-way cut, with the $r - 1$ nodes in V_s and the rest of the graph nodes forming the r components. There are $\binom{n}{r-1}$ permutations of vertices and therefore possible V_s and likewise there are $\binom{n-2}{r-1}$ possible permutations that a specific edge is not chosen (in more detail, the number of permutations removing two nodes, the nodes of the edge), so that mean that the probability of an edge not being part of this cut is:

$$\mathbb{P}_{e!\notin C} = \frac{\binom{n-2}{r-1}}{\binom{n}{r-1}} = \frac{\frac{(n-2)!}{(r-1)!(n-r-1)!}}{\frac{n!}{(r-1)!(n-r+1)!}} = \frac{(n-r+1)(n-r)}{n(n-1)}$$

That will be a lower bound for the min-cut. A simple explanation is that this specific instance is for an arbitrary cut with $|C| \leq k$, so the probability of an edge not belonging to a smaller set will be even bigger. Therefore, we have a bound for the probability of success, when we choose two of n nodes to contract. So:

$$p = \mathbb{P}_{\text{success}} = \mathbb{P}_{\text{not contracting an e}\in\mathbb{C}} = \mathbb{P}_{\text{not contracting in 1}^{\text{st}}\text{step}} \times \mathbb{P}_{\text{not contracting in 2}^{\text{nd}}\text{step}} \times \cdots \times \mathbb{P}_{\text{not contracting in n-r}^{\text{th}}\text{ step}} \geq \prod_{i=0}^{n-r-1} \frac{(n-r-1-i)(n-r-i)}{(n-i)(n-i-1)} = r\frac{1}{\binom{n}{r-1}}\frac{1}{\binom{n-1}{r-1}}$$

After determining the probability of success, presumably the rest of the analysis is the same as the 2-way Min-Cut algorithm.

Exercise 2

We will assume that we have a population A, with $|A| = n$, of which we have a $X \subseteq A$ that is in favor of the change. That means that

$$p = \frac{|X|}{n}$$

is the percentage of the whole population in favor of the change. Now, we choose another subset U, that is the people we ask for the gallop. So we have that

$$\hat{p} = \frac{|U \cap X|}{|U|}$$

is an approximation of p, based on the population that we had access to. U can be thought of as the sum of $|U|$independent variables,

$$U_i = \begin{cases} 1, & p\ (i \to \mathrm{yes}) \\ 0, & 1-p\ (i \to \mathrm{no}) \end{cases}$$

We have that:

$$\mathbb{P}(|\hat{p} - p| \leq \varepsilon p) > 1 - \delta \Rightarrow$$
$$\mathbb{P}(|n\hat{p} - np| \leq \varepsilon np) > 1 - \delta \Rightarrow$$
$$\mathbb{P}(|n\hat{p} - np| \geq \varepsilon np) \leq \delta$$

Now we have a form where the Chernoff bound* is applicable, since

$$\mu = \mathbb{E}[U] = \mathbb{E}[\sum U_i] = \sum \mathbb{E}[U_i] = np$$

So, we have:

$$\mathbb{P}(|n\hat{p} - np| \geq \varepsilon np) \leq 2e^{-\varepsilon^2 np/3} \leq \delta \Rightarrow$$
$$n \geq \frac{3 \ln \frac{2}{\delta}}{p\varepsilon^2}$$

For $\varepsilon = 0.02$ and $\delta = 0.05$ and $p \in [0.1, 0.7]$, this becomes:

$$n \geq \frac{3 \ln \frac{2}{0.02}}{0.1 \times 0.05^2} \simeq 55.262 \ people$$

With the exact same logic, we can also end up with:

$$\mathbb{P}(|n\hat{p} - np| \geq \varepsilon n) \leq \delta$$

And now, using the Hoeffding bound*, we can see that:

$$\mathbb{P}(|n\hat{p} - np| \geq \varepsilon n) \leq 2e^{-2\varepsilon^2 n} \leq \delta \Rightarrow$$
$$n \geq \frac{\ln \frac{2}{\delta}}{2\varepsilon^2}$$

So then we have that:

$$n \geq \frac{\ln \frac{2}{0.02}}{2 \times 0.05^2} \simeq 921 \ people$$

We can see that the minimum number of people that we have to ask is independent of the total population in both cases.

* Chernoff bound: $\Pr(|X - \mu| \geq \delta\mu) \leq 2e^{-\delta^2 \mu/3}, \qquad 0 < \delta < 1.$

* Hoeffding bound: $\mathbf{P}(|S_n - \mathbf{E}[S_n]| \geq t) \leq 2\exp(-2t^2/n)$, if S_n sum of Bernoulli random variables

Exercise 3

a) We will attempt to prove this using the probabilistic proof method. Specifically, we are going to show that there is a non-zero probability of $|\mathbf{a} \cdot \mathbf{x} - \mathbf{a} \cdot \mathbf{y}| \leq \varepsilon$, meaning:

$$\mathbb{P}(|\mathbf{a} \cdot \mathbf{x} - \mathbf{a} \cdot \mathbf{y}| \leq \varepsilon) > \mathbf{0}$$

We will create a random variable:

$$X_j = a_i, \ x_i, \ \forall i$$

Then we have the random variable

$$S = \frac{1}{k}\sum_{i}^{k} X_i$$

Now let's suppose we select a y so that we can write $y_i = 1/k$, and so we have that

$$S = \frac{1}{k}\sum_{i}^{k} a_i = a \cdot y$$

Then

$$\mathbb{E}[S] = \frac{1}{k}\sum_{j}^{k}\sum_{i}^{k} a_i \cdot x_i = \frac{1}{k}\sum_{i}^{k} k \cdot a_i \cdot x_i = a \cdot x$$

So we have shown that $|a \cdot x - a \cdot y| = |S - \mathbb{E}[S]|$

So now we can apply the Hoeffding bound and we are asking that:

$$\mathbb{P}(\frac{1}{k}\sum_{i}^{k} X_i - \mathbb{E}[\frac{1}{k}\sum_{i}^{k} X_i] > \varepsilon) \leq 2e^{-2k\varepsilon^2} \Leftrightarrow 1 - \mathbb{P}(|a \cdot x - a \cdot y| \leq \varepsilon) \geq 1 - 2e^{-2ke^2} > 0 \Leftrightarrow -2k\varepsilon^2 < \ln(\frac{1}{2}) \Leftrightarrow$$

$$k > \ln(2)/2\varepsilon^2$$

So we need a k bigger than the one above in order to bound our probability to be strictly greater than 0 and therefore to prove our hypothesis. Since we have:

$$k(\varepsilon) = \lfloor \ln(2)/(2\varepsilon^2) \rfloor + 1,$$

it satisfies this constraint and so we have proven the hypothesis.

b) We assume that $A = (\mathbf{a_1}, \mathbf{a_2}, ..., \mathbf{a_n})$. Then we have that:

$$\| A \cdot x - A \cdot y \|_\infty = max_i(|a_i \cdot x - a_i \cdot y|)$$

Therefore:

$$\mathbb{P}(\| A \cdot x - A \cdot y \|_\infty \geq \varepsilon) = \mathbb{P}(max_i(|a_i \cdot x - a_i \cdot y| \geq \varepsilon)) \leq \sum_{i}^{m} \mathbb{P}(|a_i \cdot x - a_i \cdot y| \geq \varepsilon)$$

using Union Bound. But from (a) that means that we must have:

$$\sum_{i}^{m} \mathbb{P}(|a_i \cdot x - a_i \cdot y| \geq \varepsilon) \leq m \times 2e^{2ne^2}$$

From (a) we had that n needed to be at least $k(\varepsilon)$ for this probability to be strictly less than 1 (meaning the 1 - probability will be strictly over 0 and therefore proving our hypothesis). So we can simply use the same logic as above (with 2m instead of 2 in the inequalities) to prove that this time we must have:

$$k(m, \varepsilon) = \left\lfloor \ln(2m)/(2\varepsilon^2) \right\rfloor + 1$$

Exercise 4

We begin by calculating the probability of an edge being part of the summation. That is:

$$\begin{aligned} &\mathbb{P}(e \in \delta(S)) = \mathbb{P}\left(u \in S_i \wedge v \in S_j \wedge i \neq j\right) \\ &= 1 - \mathbb{P}\left(u \in S_i \wedge v \in S_i\right) \end{aligned}$$

To calculate this probability, we have that, since the events of two vertices being a specific S_i are separate and their union equals with the event in the probability above:

$$\begin{aligned} &\mathbb{P}(e \in \delta(S))) = 1 - \mathbb{P}(u, v \in S_1) - \mathbb{P}(u, v \in S_2) - ... - \mathbb{P}(u, v \in S_k) \overset{\text{i.i.d.}}{=} \\ &1 - \mathbb{P}(u \in S_1)\mathbb{P}(v \in S_1) - ... - \mathbb{P}(u \in S_k)\mathbb{P}(v \in S_k) = \\ &1 - \sum_{i=1}^{k} (\frac{c_i}{|V|})^2 = p \end{aligned}$$

So, by writing the sum of weights as a sum of Bernoulli distributions:

$$W(e) = \begin{cases} w(e), & p\ (e \in \delta(S)) \\ 0, & 1 - p\ (else) \end{cases}$$

we have that:

$$\mathbb{E}[w(S)] = \mathbb{E}[\sum_{e \in \delta(S)} w(e)] = \sum_{e \in E} \mathbb{E}[W(e)] = \sum_{e \in E} (w(e) \times p) = p \sum_{e \in E} w(e)$$

We can see that this **mean is maximized** when the probability p is also maximized, and this occurs when the sum in the probability is minimized. That this is achieved when

$$c_i = c = |V|/k\ \forall i$$

meaning when the cuts are of the same size. Although intuitive, the exact reason why

this happens can be shown using the quadratic mean - algebraic mean inequality, stating that:

$$\sqrt{\frac{a_1^2 + a_2^2 + \cdots + a_n^2}{n}} \geq \frac{a_1 + a_2 + \cdots + a_n}{n}$$

$$\text{with equality if and only if} \quad a_1 = a_2 = \cdots = a_n$$

and from that we can easily derive that

$$x_1^2 + x_2^2 + \cdots + x_n^2 \geq \frac{c^2}{n}$$

so the min of the summation is c^2/n, when all cuts are equal.

Now, we aim to bound the size of the cuts with high probability. We will use Chernoff bounds:

$$\mathbb{P}(C < (1-\varepsilon)c_i) \leq e^{-\varepsilon^2 c_i/3}$$
$$\mathbb{P}(C > (1+\varepsilon)c_i) \leq e^{-\varepsilon^2 c_i/2}$$

Meaning that we can confidently say (probability decreases exponentially with the size c_i) that the size of the cuts concentrated its mean (upper and lower bound)

Naturally, a more sophisticated analysis would be the one using balls and bins, since the problem can be interpreted like this but with different probabilities.

Exercise 5

Let's assume that there is a hypothesis $h_i \in S_t$ that has a loss of $L_{\mathbb{D},f}(h_i) \geq \epsilon$.
We want to show that if for $\Omega(\log(1/\delta)/\epsilon)$ consecutive samples the hypothesis class S_t remains the same, meaning that halving algorithm makes no mistake, then each expert in the hypothesis class, h_i, satisfies the condition of PAC learning, meaning that

$$\mathbb{P}(L_{\mathbb{D},f}(h_i) \leq \epsilon) \geq 1 - \delta, \ \forall h_i \in S_t,$$

$$\text{where } L_{\mathbb{D},f}(h_i) = \mathbb{P}_{x \sim \mathbb{D}}(f(x) \neq h_i(x))$$

Since the hypothesis h_i is consistent, it means that it makes no mistakes for these m consecutive samples, meaning that the probability of it not making any mistakes in these samples it is $(1 - L_{\mathbb{D},f}(h_i))$

In the worst case, that happens $|H|$ times, because, we do $m - 1$ samples where all hypothesis are consistent and then a hypothesis is wrong, reducing the S_t. Then, the sample count resets, and this can continue until we have one last hypothesis, the one that has to be by definition the realizable one (again, this is a worst case analysis). We therefore have that:

$$\mathbb{P}(L_{\mathbb{D},f}(h_i) \geq \varepsilon | h_i \in S_t) \leq |H| \times (1-\varepsilon)^m = |H| \times (1-\varepsilon)^{\Omega(\ln(1/\delta)/\varepsilon)} \leq$$
$$|H| \times (1-\varepsilon)^{\ln(|H|/\delta)/\varepsilon)} \leq |H| \times e^{-\varepsilon \cdot \ln(|H|/\delta)/\varepsilon} = |H| \times e^{-\ln(|H|/\delta)} = |H| \frac{\delta}{|H|} = \delta$$

from which of course is equivalent to the PAC definition above.

To find the sample complexity of the algorithm, we will take into account the worst case; that is that the algorithm performs no mistake for the minimum amount before the early stopping happens. Now, we have that this happens every $\log(1/\delta)/\epsilon$ samples at minimum, meaning that we can stop the execution at $\Theta(\log(1/\delta)/\epsilon)$, and S_t can change at most $|H|$ times, which is the number of mistakes the algorithm makes at most, so we end up with a sample complexity of

$$O(|H| \log(|H|/\delta)/\epsilon)$$

Exercise 6

We will be analyzing the WMA algorithm, in the case that after observing the feedback, we decay the weights of incorrect experts with a factor of $(1-\epsilon)$instead of ½ . This is described mathematically as:

$$\forall h \in H \; with \; h(x_t) \neq y_t, \; \mathbf{w_{t+1}} = (\mathbf{1} - \epsilon)\mathbf{w_t}(\mathbf{h})$$

Therefore, every time a prediction is made, we have that:

$$W_{t+1} = \sum_{a=1}^{K} w_{t+1}(a) = \sum_{\text{wrong}} (1-\epsilon) w_t(a) + \sum_{\text{correct}} w_t(a)$$
$$= W_t - \epsilon \sum_{\text{wrong}} w_t(a) \leq W_t - \epsilon(\frac{1}{2} W_t)$$
$$= \frac{1-\epsilon}{2} W_t$$

The inequality comes from the fact that since we choose predictions based on the majority, the correct weights will always be more than half of the sum of all the weights (in binary classification tasks of course).

Let L be the number of mistakes of the best expert and M the number of mistakes of the algorithm. We have that:

$$(1-\epsilon)^L \leq |H|(1-\frac{\epsilon}{2})^M \Rightarrow L\log(1-\epsilon) \leq \log(|H|) + M\log(1-\frac{\epsilon}{2}) \xrightarrow{\log(1+x)\leq x}$$
$$M \leq \frac{2}{\epsilon}\log(|H|) + \frac{2}{1-\epsilon} L$$

This inequality above is the upper bound for the mistakes our model can have.

Exercise 7

a) We have that

$$\Phi(T) = \sum_{i=1}^{n} w_T(i) = \sum_{i=1}^{n} e^{-\eta l_t(i)} =$$
$$= e^{-\eta lt(*i)} + \sum_{\substack{i=1 \\ i \neq *i}}^{n} e^{-nlt(i)} \geqslant e^{-nl_t(*_i)}$$

since $\sum e^{-nlt(i)} \geqslant 0$ as a sum of exponentials.

b) We have:

$$\sum_i^n w_{t+1}(i) = \sum_i^n w_t(i)e^{-\eta l_t(i)} = \sum_i^\eta \left(\sum_i^n w_t(i)\right) x_t(i)e^{-\eta lt(i)} =$$
$$\sum_i^\eta w_t(i)\left(\sum_i^n x_t(i)e^{-\eta l_t(i))}\right) = \quad \phi(t)\left(\sum_t^n (i)e^{-nl_t(i)}\right)$$

So it suffices to show that:

$$\sum_i^n x_t(i)e^{-\eta l_t(i)} \leq e^{-\eta \sum_i^n x_t(i)l_t(i)\big)} \cdot e^{\eta^2 \sum_i^n x_t(i)l_t^2(i)}$$

It is:

$$\sum_i^n x_t(i)e^{-\eta l_t(i)} \leq \sum_i^n x_t(i)(1-\eta l_t(i)+\eta^2 l_t^2(i)) = \sum_i^n x_t(i) + \sum_i^n(-\eta x_t(i)l_t(i)+\eta^2 x_t(i)l_t(i)) =$$
$$1+\sum_i^n(-\eta x_t(i)l_t(i)+\eta^2 x_t(i)l_t(i)) \leq e^{\sum_i^n(-\eta x_t(i)l_t(i)+\eta^2 x_t(i)l_t(i))} = e^{-\eta\ x\cdot l_t+\eta^2 x\cdot l_t^2}\ \text{q.e.d.}$$
$$\sum_i^n x_t(i)e^{-\eta l_t(i)} \leq \sum_i^n x_t(i)(1-\eta l_t(i)+\eta^2 l_t^2(i)) = \sum_i^n x_t(i) + \sum_i^n(-\eta x_t(i)l_t(i)+\eta^2 x_t(i)l_t(i)) =$$
$$1+\sum_i^n(-\eta x_t(i)l_t(i)+\eta^2 x_t(i)l_t(i)) \leq e^{\sum_i^n(-\eta x_t(i)l_t(i)+\eta^2 x_t(i)l_t(i))} = e^{-\eta\ x\cdot l_t+\eta^2 x\cdot l_t^2}\ \text{q.e.d.}$$

Here we used these two inequalities,

$1+x \leq e^x$ and $e^x \leq 1+x+x^2 \quad$ for $x < 1.79$.

Then, since $x_i, l_{t_i}, l_{t_i}^2 \leq 1$, we have that:

$$\Phi(T) \leq \Phi(1)e^{\sum_{t=1}^{T-1} -\eta x_t\cdot l_t+\eta^2 x_t\cdot l_t^2} = ne^{\sum_{t=1}^{T-1} -\eta x_t\cdot l_t+\eta^2 x_t\cdot l_t^2} \Rightarrow$$
$$\Phi(T) \leq ne^{\eta^2 T-\eta T}$$

c) From (a) we have that:

$$\Phi(T) \geq e^{-\eta \sum_{t=1}^{T} l_t(*i)} \Rightarrow \frac{\ln(\Phi(T))}{\eta} \geq -\sum_{t=1}^{T} l_t(*i)$$

Also, since $0 \leq l_t \leq 1$ it is true that

$$\sum_{t=1}^{T} \mathbb{E}[l_t(i_t)] \leq T$$

So

$$\sum_{t=1}^{n} \mathbb{E}[l_t(i_t)] - \sum_{t=1}^{T} l_t(*i) \leq T + \frac{\ln(\Phi(T))}{\eta}$$

Now, from (b) we have that $\Phi(T) \leq n e^{\eta^2 T - \eta T} \Rightarrow \ln \Phi(T) \leq \ln(n) + \eta^2 T - \eta T$

And therefore

$$T + \frac{\ln(\Phi(T))}{\eta} \leq T + \eta T + \ln(n)\eta - T = \eta T + \frac{\ln(n)}{\eta}$$

So we end up with:

$$\sum_{t=1}^{n} \mathbb{E}[l_t(i_t)] - \sum_{t=1}^{T} l_t(*i) \leq \eta T + \frac{\ln(n)}{\eta}$$

To see when this is minimized, we will choose the min upper bound, meaning the one that has a interval equal to zero:

$$f(\eta) = \eta T + \frac{\ln(n)}{\eta} \Rightarrow f'(\eta) = T - \frac{\ln(n)}{\eta^2} = 0 \Rightarrow \eta = \sqrt{\frac{ln(n)}{T}}$$

For this *η* we achieve an Exp-Loss of:

$$\sqrt{\frac{ln(n)}{T}} T + \frac{\ln(n)}{\sqrt{\frac{ln(n)}{T}}} = 2\sqrt{ln(n)T}$$