NAME - ANNU, ROLL.NO. - EE21RESCH01010

SOURCE CODING ASSIGNMENT 1

Collaborated with - Shantanu Yadav, EE20MTECH12001

Books referred - Thomas & Cover on Information Theory

Exercise 1.1 - part 1 - prove that

$$\lim_{n\to\infty} Pr\left[(x^n, y^n) \notin T_\varepsilon^{(n)}(P_{xy})\right] = 0$$

Solution - Let (x^n, y^n) be sequences of length n i.i.d according to

$$p(x^n, y^n) = \prod_{i=1}^{n} p(x_i, y_i)$$

$$T_\varepsilon^{(n)}(P_{xy}) = \left\{ \underset{\in \mathcal{X}^n \times \mathcal{Y}^n}{(x^n, y^n)} : \left| \frac{\mu_{x^n y^n}(a,b)}{n} - P_{xy}(a,b) \right| \leq \varepsilon P_{xy}(a,b) \ \forall\ a, b \in \mathcal{X}, \mathcal{Y} \right\}$$

Type : $\dfrac{\mu_{x^n, y^n}(a,b)}{n} = \dfrac{\#\text{ of } a\text{'s in } x^n,\ b\text{'s in } y^n}{n}$, $a \in \mathcal{X}$, $b \in \mathcal{Y}$

$= \#(a,b)$'s in $x^n \times y^n$

Let indicator function

$$1_{a,b} = \begin{cases} 1, & (a,b) = (x,y) \\ 0, & \text{otherwise} \end{cases} \quad ; \quad \mu_{x^n, y^n}(a,b) = \sum_{i=1}^{n} 1_{a,b}(x_i, y_i)$$

$$\therefore \frac{\mu_{x^n, y^n}(a,b)}{n} = \frac{\sum_{i=1}^{n} 1_{a,b}(x_i, y_i)}{n} = \text{type} \left\{ \because \mu_{x^n y^n}(a,b) = \#(a,b)\text{'s in } x^n \times y^n \right\}$$

$$E\frac{\mu_{x^n, y^n}(a,b)}{n} = \sum_{i=1}^{n} \frac{E\, 1_{a,b}(x_i, y_i)}{n} = P_{xy}(a,b) = \text{true joint pmf for } (a,b) \in \mathcal{X} \times \mathcal{Y}$$

Now, consider - $Pr\left[(x^n, y^n) \notin T_\varepsilon^{n}(P_{xy})\right]$

$$= Pr\left[\left| \frac{\mu_{x^n y^n}(a,b)}{n} - P_{xy}(a,b) \right| > \varepsilon P_{xy}(a,b) \text{ for some } a,b \in \mathcal{X} \times \mathcal{Y} \right]$$

$$\leq \sum_{(a,b) \in \mathcal{X} \times \mathcal{Y}} Pr\left[\left| \frac{\sum_{i=1}^{n} 1_{a,b}(x_i, y_i)}{n} - P_{xy}(a,b) \right| > \varepsilon P_{x,y}(a,b) \right]$$

Scanned with CamScanner

Pr

Using chernoff's bound -

$$\Pr\left[(X^n, Y^n) \notin T_\varepsilon^{(n)}(P_{XY}(a,b))\right] \leq \sum_{(a,b)\in \mathcal{X}\times\mathcal{Y}} 2 \times e^{-n\varepsilon^2 P_{XY}(a,b)/3}$$

$$\leq 2^{-n\varepsilon^2 \alpha(P_{XY})}$$ for some α which depends on P_{XY}.

$$\Rightarrow \Pr\left[(X^n, Y^n) \notin T_\varepsilon^{(n)}(P_{XY}(a,b))\right] \leq 2^{-n\varepsilon^2\alpha}$$
for some $(a,b) \in \mathcal{X}\times\mathcal{Y}$

As $n \to \infty$, $2^{-n\varepsilon^2\alpha} \to 0$.

$\therefore$ As $n \to \infty$

$$\Pr\left[(X^n, Y^n) \notin T_\varepsilon^{(n)}(P_{XY})\right] \leq 2^{-n\varepsilon^2\alpha}$$

$$\Pr\left[(X^n, Y^n) \notin T_\varepsilon^{(n)}(P_{XY})\right] \leq 0$$

And since probability is non-negative & so,

$$\boxed{\text{As } n \to \infty \quad \Pr\left[(X^n, Y^n) \notin T_\varepsilon^{(n)}(P_{XY})\right] \to 0}$$

for sufficiently large n. Hence proved

Scanned with CamScanner

Exercise 1.1 - part 2 -

$\forall\ g: \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ and all $(x^n, y^n) \in T_\varepsilon^{(n)}(P_{XY})$

$g(x, y) = g(x_i, y_i)\ \forall\ (x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$

consider.

$$\sum_{i=1}^{n} g(x_i, y_i) = \sum_{i=1}^{n} \sum_{(a,b) \in (\mathcal{X} \times \mathcal{Y})} 1_{(a,b)}(x_i, y_i)\, g(a, b) \quad \text{--- (1)}$$

where $1_{a,b}(x, y) = \begin{cases} 1, & (a,b) = (x, y) \\ 0, & (a,b) \neq (x, y) \end{cases}$

Indicator function

from eqn (1) -

$$\boxed{\sum_{i=1}^{n} g(x_i, y_i) = \sum_{(a,b) \in \mathcal{X} \times \mathcal{Y}} \left(\sum_{i=1}^{n} 1_{a,b}(x_i, y_i) \right) g(a, b)} \quad \text{--- (A)}$$

Now, from definition of jointly typical sets -

$$T_\varepsilon^{(n)}(P_{XY}) = \{ (x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : |\mu_{x^n, y^n}(a,b)/n - P_{XY}(a,b)| \leq \varepsilon P_{XY}(a,b)\ \forall\ (a,b) \in \mathcal{X} \times \mathcal{Y} \}$$

where $\mu_{x^n, y^n}(a, b)$ is the locations (no. of locations) for which $i \in \{1, 2, \ldots n\}$ & $x_i = a$ & $y_i = b$.

i.e type: $\dfrac{\mu_{x^n, y^n}(a,b)}{n} = \dfrac{\text{no. of } (a,b)\text{'s in } x^n \times y^n}{n}$, $(a,b) \in \mathcal{X} \times \mathcal{Y}$

Using Indicator function

$$\frac{\mu_{(x^n, y^n)}(a, b)}{n} = \frac{\sum_{i=1}^{n} 1_{a,b}(x_i, y_i)}{n}$$

where $1_{a,b}(x, y) = \begin{cases} 1, & (a,b) = (x, y) \\ 0, & (a,b) \neq (x, y) \end{cases}$

$$\frac{E\, \mu_{(x^n, y^n)}(a, b)}{n} = \sum_{i=1}^{n} \frac{E\, 1_{a,b}(x_i, y_i)}{n} = P_{X,Y}(a,b) = \text{joint Prob}$$

Scanned with CamScanner

Now, as consequence of chernoff's bounds -

$$\Pr\left[\,|\mu_{x^n,y^n}(a,b)|n - P_{x,y}(a,b)| > \varepsilon\, P_{xy}(a,b) \text{ for some } (a,b) \in \mathcal{X}\times\mathcal{Y}\right]$$

$$\leq \sum_{(a,b)\in\mathcal{X}\times\mathcal{Y}} \Pr\left[\,|\mu_{x^n,y^n}(a,b)|n - P_{x,y}(a,b)| > \varepsilon\, P_{x,y}(a,b)\right]$$

$$= \Pr\left[\left|\frac{\sum_{i=1}^{n} 1_{a,b}(x_i,y_i)}{n} - P_{x,y}(a,b)\right| > \varepsilon\, P_{x,y}(a,b)\right] \quad - (2)$$

(Using chernoff's bound)

$$\leq \sum_{(a,b)\in\mathcal{X}\times\mathcal{Y}} 2\times e^{-n\varepsilon^2 P_{xy}(a,b)/3} \leq 2^{-n\varepsilon^2\alpha(P_{xy})}$$

for some α, which depends on P_{xy}

consider eqn (2)

$$\Pr\left[\left|\frac{\sum_{i=1}^{n} 1_{a,b}(x_i,y_i)}{n} - P_{xy}(a,b)\right| > \varepsilon\, P_{xy}(a,b)\right]$$

$$\leq \sum_{(a,b)\in\mathcal{X}\times\mathcal{Y}} 2e^{-n\varepsilon^2 \frac{P_{xy}(a,b)}{3}}$$

$$\Rightarrow \Pr\left[x^n,y^n \in T_\varepsilon^{(n)}(P_{xy})\right] \geq 1 - 2^{n\varepsilon^2\alpha(P_{xy})}$$

$$\Rightarrow \left|\frac{\sum_{i=1}^{n} 1_{a,b}(x_i,y_i)}{n} - P_{xy}(a,b)\right| \leq \varepsilon\, P_{xy}(a,b)$$

$$\Rightarrow \frac{\sum_{i=1}^{n} 1_{a,b}(x_i,y_i)}{n} - P_{xy}(a,b) \leq \varepsilon\, P_{xy}(a,b) \quad - (3)$$

$$\& \quad \frac{\sum_{i=1}^{n} 1_{a,b}(x_i,y_i)}{n} - P_{xy}(a,b) \geq -\varepsilon\, P_{xy}(a,b) \quad - (4)$$

from eqn (4) -

$$\boxed{\frac{\sum_{i=1}^{n} 1_{a,b}(x_i,y_i)}{n} \geq (1-\varepsilon)\, P_{xy}(a,b)} \quad - (5)$$

from eqn (3) -

$$\boxed{\frac{\sum 1_{a,b}(x_i,y_i)}{n} \leq (1+\varepsilon)\, P_{x,y}(a,b)} \quad - (6)$$

Scanned with CamScanner

Now from eqn (A) -

$$\sum_{i=1}^{n} g(x_i, y_i) = \sum_{(a,b)\in \mathcal{X}\times\mathcal{Y}} \left(\sum_{i=1}^{n} I_{a,b}(x_i, y_i) \right) g(a,b)$$

Using eqn (5) for upper bound -

eqn (A) becomes -

$$\sum_{i=1}^{n} g(x_i, y_i) = \sum_{(a,b)\in \mathcal{X}\times\mathcal{Y}} \left(\sum_{i=1}^{n} I_{a,b}(x_i, y_i) \right) g(a,b)$$

$$\geq \sum_{(a,b)\in \mathcal{X}\times\mathcal{Y}} n(1-\varepsilon)\, P_{XY}(a,b)\, g(a,b)$$

$$= n(1-\varepsilon) \sum_{(a,b)\in \mathcal{X}\times\mathcal{Y}} P_{XY}(a,b)\, g(a,b)$$

$$\Rightarrow \boxed{\sum_{i=1}^{n} g(x_i, y_i) \geq n(1-\varepsilon) \sum_{(a,b)\in \mathcal{X}\times\mathcal{Y}} P_{XY}(a,b)\, g(a,b)} \quad -(B)$$

Now, using eqn (6) for lower bound -

eqn (A) becomes -

$$\sum_{i=1}^{n} g(x_i, y_i) = \sum_{(a,b)\in \mathcal{X}\times\mathcal{Y}} \left(\sum_{i=1}^{n} I_{a,b}(x_i, y_i) \right) g(a,b)$$

$$\leq \sum_{(a,b)\in \mathcal{X}\times\mathcal{Y}} n(1+\varepsilon)\, P_{XY}(a,b)\, g(a,b)$$

$$= n(1+\varepsilon) \sum_{(a,b)\in \mathcal{X}\times\mathcal{Y}} P_{XY}(a,b)\, g(a,b)$$

$$\Rightarrow \boxed{\sum_{i=1}^{n} g(x_i, y_i) \leq n(1+\varepsilon) \sum_{(a,b)\in \mathcal{X}\times\mathcal{Y}} P_{XY}(a,b)\, g(a,b)} \quad -(C)$$

But as we know -

$$\sum_{(a,b)\in \mathcal{X}\times\mathcal{Y}} P_{X,Y}(a,b)\, g(a,b) = E[g(X,Y)] \quad -(D)$$

Scanned with CamScanner

feeding value of $\sum_{(a,b)\in X\times Y} P_{XY}(a,b)\, g(a,b) = E[g(x,y)]$ in

eqn (B) & (C), we get -

from (B) -

$$\sum_{i=1}^{n} g(x_i, y_i) \geq n(1-\varepsilon)\, E[g(x,y)]$$

$$\Rightarrow \frac{1}{n}\sum_{i=1}^{n} g(x_i, y_i) \geq (1-\varepsilon)\, E[g(x,y)] \quad -(\mathrm{I})$$

Similarly from (C) -

$$\sum_{i=1}^{n} g(x_i, y_i) \leq n(1+\varepsilon)\, E[g(x,y)]$$

$$\Rightarrow \frac{1}{n}\sum_{i=1}^{n} g(x_i, y_i) \leq (1+\varepsilon)\, E[g(x,y)] \quad -(\mathrm{II})$$

Combining eqn (I) & (II), we get -

$$\boxed{(1-\varepsilon)\, E[g(x,y)] \leq \frac{1}{n}\sum_{i=1}^{n} g(x_i, y_i) \leq (1+\varepsilon)\, E[g(x,y)]}$$

QED//

Scanned with CamScanner

Exercise 1.1 - part 3 -

Bounds on $|T_\varepsilon^{(n)}| = ?$

Expression obtained in part 2 Exercise 1.1 is -

$$(1-\varepsilon)E[g(X,Y)] \leq \frac{1}{n}\sum_{i=1}^{n} g(x_i, y_i) \leq (1+\varepsilon)E[g(X,Y)] \quad -(1)$$

when $g(x_i, y_i) = \log_2 \frac{1}{p(x_i, y_i)}$

or $g(x,y) = \log_2 \frac{1}{p(x,y)}$

equation (1) becomes -

$$(1-\varepsilon)E\left[\log_2 \frac{1}{p(x,y)}\right] \leq \frac{1}{n}\sum_{i=1}^{n} \log_2 \frac{1}{p(x_i,y_i)} \leq (1+\varepsilon)E\left[\log_2 \frac{1}{p(x,y)}\right] \quad -(2)$$

consider $\frac{1}{n}\sum_{i=1}^{n} \log_2 \frac{1}{p_{X,Y}(x_i,y_i)} = \frac{1}{n}\log \frac{1}{\prod_{i=1}^{n} p_{X,Y}(x_i,y_i)}$

$$= \frac{1}{n}\log \frac{1}{p_{X^n,Y^n}(x^n,y^n)} \quad -(3)$$

{ $\because$ x^n, y^n is Memoryless sequence }

$\therefore$ eqn (2) becomes -

$$(1-\varepsilon)\underbrace{E\left[\log_2 \frac{1}{p_{X,Y}(x,y)}\right]}_{H(X,Y)} \leq \frac{1}{n}\sum_{i=1}^{n} \log_2 \frac{1}{p_{X,Y}(x_i,y_i)} \leq (1+\varepsilon)\underbrace{E\left[\log_2 \frac{1}{p_{X,Y}(x,y)}\right]}_{H(X,Y)}$$

$\Rightarrow$

$$\Rightarrow (1-\varepsilon)H(X,Y) \leq \frac{1}{n}\log \frac{1}{p_{X^n,Y^n}(x^n,y^n)} \leq (1+\varepsilon)H(X,Y)$$

(using (3))

$$\Rightarrow \boxed{2^{-n(1+\varepsilon)H(X,Y)} \leq p_{X^n,Y^n}(x^n,y^n) \leq 2^{-n(1-\varepsilon)H(X,Y)}}$$

(taking inverse log)

Splitting above eqn into 2 parts -

$$p_{X^n,Y^n}(x^n,y^n) \leq 2^{-n(1-\varepsilon)H(X,Y)} \quad -(4)$$

$$p_{X^n,Y^n}(x^n,y^n) \geq 2^{-n(1+\varepsilon)H(X,Y)} \quad -(5)$$

Scanned with CamScanner

Consider - $\Pr[(X^n,Y^n) \in T_\varepsilon^{(n)}(P_{XY})] \leq 1$

$1-2^{-n\varepsilon^2\alpha} \leq \Pr[(X^n,Y^n) \in T_\varepsilon^{(n)}(P_{XY})] \leq 1$ —(a) {since probability of Joint typical set is less than or equal to 1}

for some α, which depends on P_{XY}.

Now from (4) -

$$\Pr[(X^n,Y^n) \in T_\varepsilon^{(n)}(P_{XY})] = \sum_{x^n,y^n \in T_\varepsilon^{(n)}(P_{XY})} P_{X^nY^n}(x^n,y^n) \leq \sum_{x^n,y^n \in T_\varepsilon(P_{XY})} 2^{-n(1-\varepsilon)H(X,Y)} \geq 1-2^{-n\varepsilon^2\alpha}$$

{using eqn (a) lower bound}

$$\Rightarrow \sum_{x^n,y^n \in T_\varepsilon^{(n)}(P_{XY})} 2^{-n(1-\varepsilon)H(X,Y)} \geq 1-2^{-n\varepsilon^2\alpha}$$

$$\Rightarrow \sum_{(x^n,y^n) \in T_\varepsilon^{(n)}(P_{XY})} 1 \geq (1-2^{-n\varepsilon^2\alpha})\, 2^{nH(X,Y)(1-\varepsilon)}$$

$$\Rightarrow |T_\varepsilon^{(n)}(P_{XY})| \geq (1-2^{-n\varepsilon^2\alpha})\, 2^{nH(X,Y)(1-\varepsilon)}$$

for some α, which depends on P_{XY}.

$$\Rightarrow \boxed{|T_\varepsilon^{(n)}(P_{XY})| \geq (1-2^{-n\varepsilon^2\alpha})\, 2^{nH(X,Y)(1-\varepsilon)}} \quad -(A)$$

Again, $\Pr[x^n,y^n \in T_\varepsilon^{(n)}(P_{XY})] \leq 1$ (from eqn (a) upper bound)

from (5)

$$\Pr[(X^n,Y^n) \in T_\varepsilon^{(n)}(P_{XY})] = \sum_{x^n,y^n \in T_\varepsilon^{(n)}(P_{XY})} P_{X^nY^n}(x^n,y^n) \leq 1$$

$$\geq \sum_{x^n,y^n \in T_\varepsilon(P_{XY})} 2^{-nH(X,Y)(1+\varepsilon)}$$

$$\Rightarrow \sum_{x^n,y^n \in T_\varepsilon(P_{XY})} P_{X^nY^n}(x^n,y^n) \geq \sum_{x^n,y^n \in T_\varepsilon(P_{XY})} 2^{-nH(X,Y)(1+\varepsilon)} \leq 1$$

$$\Rightarrow \sum_{(x^n,y^n) \in T_\varepsilon(P_{X,Y})} 2^{-nH(X,Y)(1+\varepsilon)} \leq 1$$

$$\Rightarrow \sum_{(x^n,y^n) \in T_\varepsilon^{(n)}(P_{XY})} \leq 2^{nH(X,Y)(1+\varepsilon)}$$

$$\Rightarrow \boxed{|T_\varepsilon^{(n)}(P_{XY})| \leq 2^{nH(X,Y)(1+\varepsilon)}} \quad -(B)$$

from (A) & (B)

$\therefore$ Final Bounds - $(1-2^{-n\varepsilon^2\alpha})\, 2^{nH(X,Y)(1-\varepsilon)} \leq |T_\varepsilon^{(n)}(P_{XY})| \leq 2^{nH(X,Y)(1+\varepsilon)}$ for some $\varepsilon > 0$, α depends on P_{XY}

Scanned with CamScanner

Also, for sufficiently large n, $\Pr\{(x^n, y^n) \in T_\varepsilon^{(n)}(P_{XY})\} > 1-\varepsilon$

So that,

$$1-\varepsilon < \Pr\{(x^n, y^n) \in T_\varepsilon^{(n)}(P_{XY})\}$$

$$\leq \sum_{x^n, y^n \in T_\varepsilon^{(n)}(P_{XY})} 2^{-n(H(X,Y)(1-\varepsilon)} \quad \text{(from eqn}$$

$$= 2^{-n(1-\varepsilon)H(X,Y)} \left| T_\varepsilon^{(n)}(P_{XY}) \right|$$

$$\Rightarrow \left| T_\varepsilon^{(n)}(P_{XY}) \right| \geq (1-\varepsilon) \cdot 2^{n(1-\varepsilon)H(X,Y)} \quad \text{— (C)}$$

Above eqn (C) is also a lower bound for $\left| T_\varepsilon^{(n)}(P_{XY}) \right|$

So, using (B) & (C), we can write –

$$\boxed{(1-\varepsilon) \cdot 2^{n(1-\varepsilon)H(X,Y)} \leq \left| T_\varepsilon^{(n)}(P_{XY}) \right| \leq 2^{nH(X,Y)(1+\varepsilon)}}$$

Bounds for $T_\varepsilon^{(n)}(P_{XY})$

Scanned with CamScanner

Problem-2 - part-2

From python code -

Empirical pmf = {'c': 0.592, 'a': 0.209, 'e': 0.034, 'b': 0.094, 'd': 0.071}

length of the file = 1000

X = symbols = {'c', 'a', 'b', 'd', 'e'}

$P(c) = 0.592$, $P(a) = 0.209$, $P(b) = 0.094$

$P(d) = 0.071$, $P(e) = 0.034$

Entropy of file

$$H(x) = +\sum P(x)\log_2 1/P(x) = 1.67721 \text{ bits}$$

Optimal compressed file size calculated using entropy = 209.65 bytes

optimal compressed file size calculated using entropy = $H(x) \times 8/1000$

Using Huffman coding

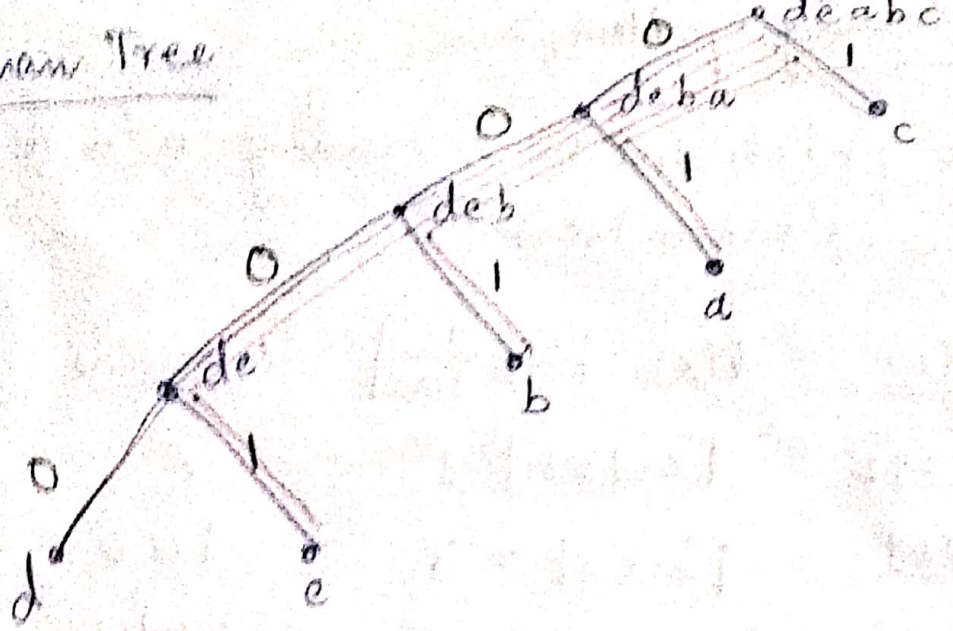
let's arrange probabilities in descending order for better formation of tree (Huffman tree)

Symbols	P(x)			
c	0.592	→ 0.592	→ 0.592	→ 0.592 → 1
a	0.209	→ 0.209	→ 0.209	0.408
b	0.094	→ 0.094		0.199
d	0.071	→ 0.105		
e	0.034			

At each succeeding step, 2 symbols with least probability are merged & it forms a new symbol consisting previous 2 symbols & this process is repeated till we get the sum of probabilities of all symbols as 1.

Scanned with CamScanner

Huffman Tree



Assigning left of tree as 0 & right as 1 to all left & right branches of Huffman tree.

So, final codewords for each symbol, we get is –

c = 1

a = 0 1

b = 0 0 1

d = 0 0 0 0

e = 0 0 0 1

length of compressed file, calculated using Huffman coding = 216 bytes.

length without compression = 1000 bytes

∴ Rate of compression using Huffman coding

$$R = \frac{216}{1000} = 0.216$$

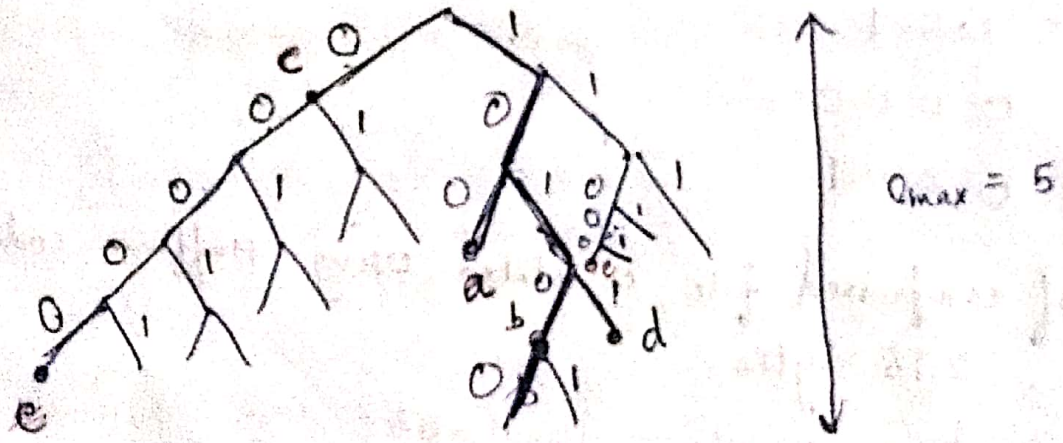
Scanned with CamScanner

Using shannon coding - problem-2

Let's arrange probabilities in descending order for best formation of tree (shannon tree).

Symbols	P(x)	$l(x) = \lceil \log_2 \frac{1}{P(x)} \rceil$	Codewords
c —	0.592	$\lceil 0.756 \rceil = 1$ —	0
a —	0.209	$\lceil 2.25 \rceil = 3$ —	100
b —	0.094	$\lceil 3.411 \rceil = 4$ —	1010
d —	0.071	$\lceil 3.81 \rceil = 4$ —	1011
e —	0.034	$\lceil 4.878 \rceil = 5$ —	11000

making a shannon tree for which maximum length = 5 ($\because l_{max} = 5$), length of each symbol $= \lceil \log_2 \frac{1}{P(x)} \rceil$



Assigning left of all branches in tree as 0 & right of all branches in tree as 1

Start from left & assign final codewords to all symbols such that no leafs got overlaps. to other leafs in tree

c = 0
a = 100
b = 1010
d = 1011
e = 11000

} final codewords

length of compressed file calculated using shanon coding

= 258. bytes

length of file w/o compression = 1000 bytes

Rate of compression (R) = $\frac{258}{1000}$

$$\boxed{R = 0.258}$$

Scanned with CamScanner

Using Shannon-fano alias coding - problem 2

Arrange probabilities in descending order.

Symbols x	$P(x)$	$\bar{F}(x)$	$\bar{F}(x)$ in binary	$l(x) = \lceil \log_2 \frac{1}{P(x)} \rceil + 1$	Codewords
c	0.592	0.296	0.01001011	2	01
a	0.209	0.6965	0.10110010	4	1011
b	0.094	0.848	0.11011100	5	11011
d	0.071	0.9305	0.11110111	5	11101
e	0.034	0.983	0.11111011	6	111110

$$\boxed{\bar{F}(x_i) = \sum_{j=1}^{i-1} P(x_j) + \frac{P(x_i)}{2}}$$

codewords are assigned to symbols using probabilities & modified CDF $\bar{F}(x_i)$

length of each codeword x -

$$l(x) = \left\lceil \log_2 \frac{1}{P(x)} \right\rceil + 1$$

So, codewords = $\bar{F}(x_i)$ in binary truncated at $l(x)$ bits

= $\bar{F}(x_i)$ in binary of length $l(x_i)$ bits

length of compressed file calculated using shannon fano alias - = 383 bytes

length w/o compression = 1000 bytes

Rate using shannon fano alias $= \frac{383}{1000}$

$$\boxed{R = 0.383}$$

Scanned with CamScanner

Problem 3 - part 2

From python code

Empirical pmf = {'e': 0.172, 'a': 0.416, 'c': 0.258, 'b': 0.153, 'd': 0.001}

Entropy of file $= + \sum p(x) \log_2 1/p(x)$

$= 1.8918$ bits

Length of file = 1000 bytes

Optimal compressed file size calculated using entropy

$= \frac{H(x) \times 1000 \text{ bytes}}{8}$

$= \cancel{209.65 \text{ bytes}}$ 236.4758 bytes

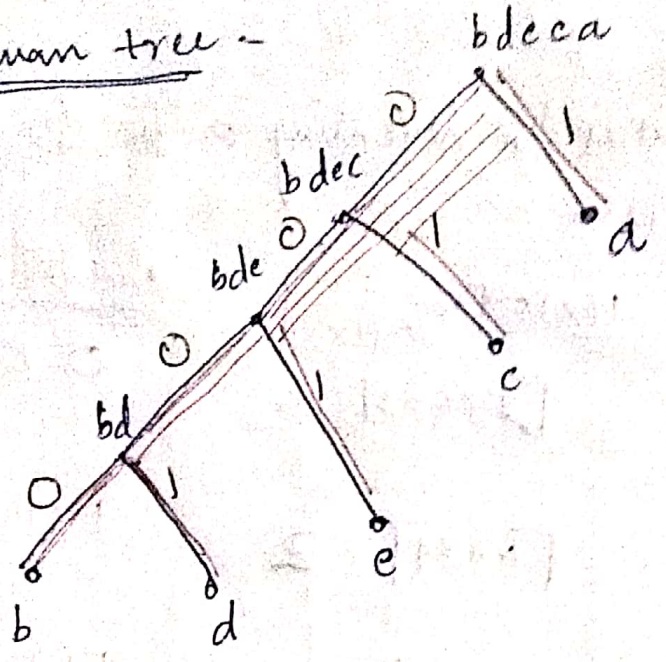
Using Huffman coding -

Let's assange probabilities in descending order for best formation of Huffman tree.

Symbols	P(x)			
a	0.416	0.416	0.416	0.416
c	0.258	0.258	0.258	0.584
e	0.172	0.172	0.326	
b	0.153	0.159		
d	0.001			

At each succeding step, 2 symbols with least probabilities are merged & it forms a new symbol consisting of 2 previous symbols & this process is repeated till we get sum of all probabilities as

Huffman tree -



Assigning left of tree as 0 & right as 1 to all left & right branches of Huffman tree.

so, final codewords are

a = 1

c = 01

e = 001

d = 0001

b = 0000

length of compressed file calculated using huffman coding = 260 bytes

length of file without compression = 1000 bytes

∴ Rate of compression using huffman coding -

$$\boxed{R = \frac{260}{1000} = 0.26}$$

Scanned with CamScanner

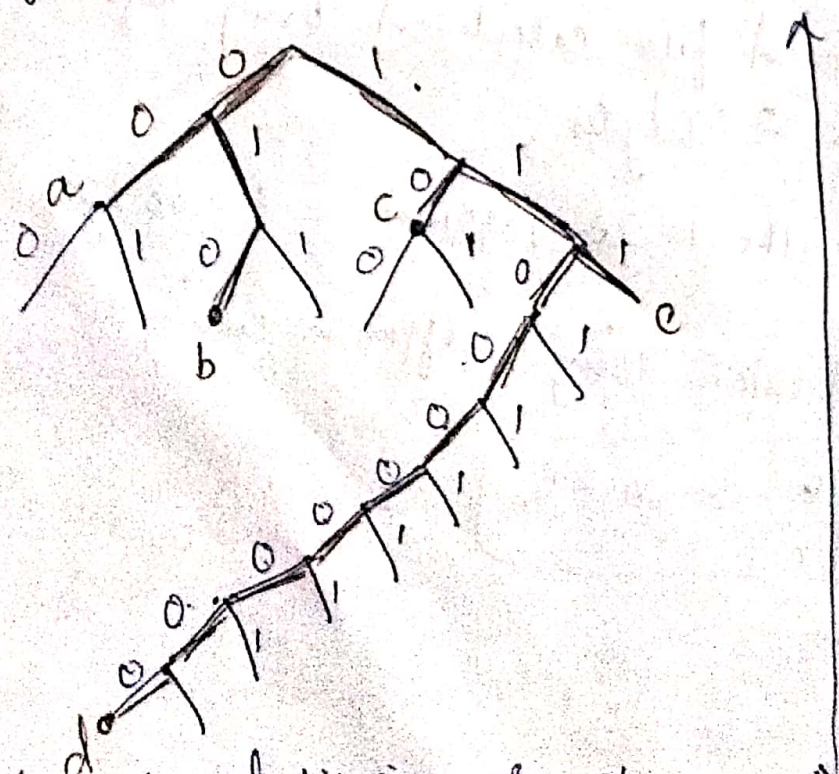
Using shannon-coding, part 2 -

Let's arrange probabilities in descending order for best formation of shannon tree.

Symbols	$P(x)$	$l(x) = \lceil \log_2 \frac{1}{P(x)} \rceil$	codewords
a	0.416	$\lceil 1.2653 \rceil = 2$	00
c	0.258	$\lceil 1.954 \rceil = 2$	10
e	0.172	$\lceil 2.539 \rceil = 3$	111
b	0.153	$\lceil 2.4087 \rceil = 3$	010
d	0.001	$\lceil 2.708 \rceil = 10$	1000000000

Making a shannon tree for which depth of tree is 10 (l_{max}). ($\because l_{max} = 10$)

length of each symbol is given by $\lceil \log_2 \frac{1}{P(x)} \rceil$



l_{max} = depth = 10

length of compressed filesize using shannon coding = 293 bytes

Assigning left of all branches in tree as 0 & right as 1 start from left & assign codewords to each symbol such that no leaf got overlap to other leaf in a tree

a = 0 0

c = 1 0

e = 1 1 1

b = 0 1 0

d = 1 1 0 0 0 0 0 0 0 0

final codewords

$$\boxed{\text{Rate} = \frac{293}{1000} = 0.293}$$

Scanned with CamScanner

Using Shannon-fano alias coding problem-3

Arrange probabilities in decending order -

Symbols x	P(x)	$\bar{F}(x)$	$\bar{F}(x)$ in binary	$l(x) = \lceil \log_2 \frac{1}{P(x)} \rceil + 1$
a	0.416	0.208	0.0011010100	3
c	0.258	0.545	0.1000101111	3
e	0.172	0.76	0.1100001101	4
b	0.153	0.9225	0.1110110000	4
d	0.001	0.9995	0.1111111111110	11

$$\boxed{\bar{F}(x_i) = \sum_{j=1}^{i-1} p(x_j) + \frac{P(x_i)}{2}}$$

Codewords are assigned to symbols using probabilities & modified CDF $\bar{F}(x_i)$

length of each codeword x -

$$l(x) = \left\lceil \log_2 \frac{1}{P(x)} \right\rceil + 1$$

So, codewords = $\bar{F}(x_i)$ in binary truncated at $l(x)$ bits

= $\bar{F}(x_i)$ in binary of length $l(x)$ bits

<u>final codewords</u> - a = 001

c = 100

e = 1100

b = 1110

d = 11111111110

length of compressed file size using Shannon-fano alias = 418 bytes

$$\boxed{\text{Rate} = \frac{418}{1000} = 0.418}$$

Scanned with CamScanner