ORIE 4742 - Info Theory and Bayesian ML

Lecture 4: Source Coding

February 4, 2020

Sid Banerjee, ORIE, Cornell

entropy and information

rv X taking values $\mathcal{X} = \{a_1, a_2, \ldots, a_k\}$, with pmf $\mathbb{P}[X = a_i] = p_i$

Shannon's entropy function

- outcome $X = a_i$ has *information content*: $h(a_i) = \log_2\left(\frac{1}{p_i}\right)$
- random variable X has *entropy*: $H(X) = \mathbb{E}[h(X)] = \sum_{i=1}^{k} p_i \log_2\left(\frac{1}{p_i}\right)$

- only depends on distribution of X (i.e., $H(X) = H(p_1, p_2, \ldots, p_k)$)
- $H(X) \geq 0$ for all X
- if $X \perp\!\!\!\perp Y$, then $H(X, Y) = H(X) + H(Y)$
  where joint entropy $H(X, Y) \triangleq \sum_{(x,y)} p(x, y) \log_2 1/p(x, y)$
- if $X \sim$ uniform on $\mathcal{X}$, then $H(X) = \log_2 |\mathcal{X}|$; else, $H(X) \leq \log_2 |\mathcal{X}|$

the source coding problem

suppose we are given a database $D = (X_1 X_2 \dots X_N)$, where each X_i is a letter in an alphabet $\mathcal{X}$, generated iid according to $X_i \sim \{p_1, p_2, \dots, p_k\}$

lossless compression

compress D into a *codeword* $L = \phi(D)$ such that can recover $D = \phi^{-1}(L)$

Shannon's source coding theorem

if X has entropy $H(X)$, then for any $\epsilon > 0$ and $0 < \delta < 1$, there exists N_0 s.t. for all $N > N_0$, we have a lossless code $L = \phi(X^N)$ s.t.

$$\left| \frac{\mathbb{E}[L]}{N} - H(X) \right| \leq \epsilon$$

Moreover, this is the best possible, i.e., no lossless code has $\frac{\mathbb{E}[L]}{N} < (1-\epsilon) H(X)$

lossless compression via typical set encoding

typical set

iid source produces $X^N = (X_1 X_2 \ldots X_N)$; each $X_i \in \mathcal{X}$ has entropy $H(X)$

then X^N is very likely to be one of $\approx 2^{NH(X)}$ typical strings,
all of which have probability $\approx 2^{NH(X)}$

Typical Set
- Size $\sim 2^{NH(x)}$
- ~Uniform distn
- Total prob $\geq 1-\delta$

(w) Size $2^{N(1-H(x))}$
prob $< \delta$

practical source coding solutions

symbol codes

$X_1 X_2 \ldots X_n \quad \rightarrow \quad \phi(X_1)\phi(X_2)\ldots\phi(X_n)$

map each symbol $X_i \rightarrow \phi(x_i)$

Eg - Morse code, Huffman code

stream codes

$X_1 X_2 \ldots X_n \quad \rightarrow \quad \phi(X_1)\phi(X_2|X_1)\phi(X_3|X_1X_2)\ldots\phi(X_n|X_1X_2\ldots X_{n-1})$

maps entire data base to a code_word

Eg - LZW
Lempel-Ziv-Welch code, arithmetic coding

symbol codes (Mackay chapter 5)

symbol codes

expected length of symbol code

let $X \sim \{p(x)\}_{x \in \mathcal{X}}$, and consider code $C(\cdot)$, and let $\ell(x) = |C(x)|$

the expected length of C is $\mathbb{E}[L(C, X)] = \sum_x p(x)\ell(x)$

what we want from symbol code C:

↷ $\forall x \neq y$, want $C(x) \neq C(y)$ (singular code)

- unique decodability: $\forall\, x_1 x_2 \ldots x_n \neq y_1 y_2 \ldots y_n$, we have $C(x_1)C(x_2)\ldots C(x_n) \neq C(y_1)C(y_2)\ldots C(y_n)$
- easy to decode
- small $\mathbb{E}[L(C, X)]$

(If sending N symbols $C(x_1)C(x_2)\ldots C(x_n)$

$\mathbb{E}\left[\sum_{i=1}^{n} |C(x_i)|\right] = \sum_{i=1}^{n} \mathbb{E}[|C(x_i)|] = N\,\mathbb{E}[L(C,X)]$)

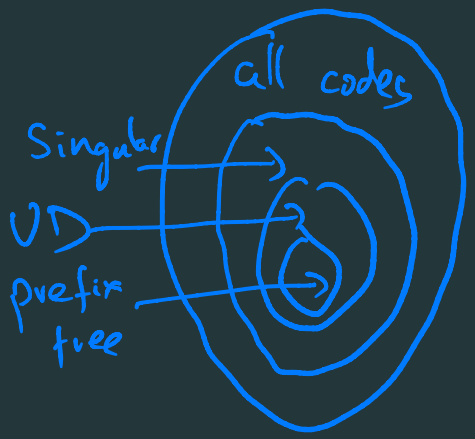
types of symbol codes

consider source producing $X \sim \{a, b, c, d\}$ with prob $\{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\}$

Symbols	info content (bits)	Code 1	$l(x)$	Code 2	$l(x)$
a	1	1000	4	00	2
b	2	0100	4	01	2
c	3	0010	4	10	2
d	3	0001	4	11	2
	$\frac{7}{4} = 1.75$ bits		4 bits on avg per symbol		2 bits

prefix codes (variable length codes, VLC)

Symbols	Info content	Code 3 (Prefix code)	Code 4 (Uniquely decodable)
a	1	0	0
b	2	10	01
c	3	110	011
d	3	111	111
	1.75	$E[L(C,x)] = 1.75$	$E[L(C,x)] = 1.75$



prefix code $\equiv$ $C(x)$ is not the 'prefix' of any $C(y)$ $\forall x, y$

(instantaneous code)

the limits of unique decodability

Kraft-McMillan inequality

for any $C \equiv$ uniquely decodable binary code over $\mathcal{X}$, with $\ell(x) = |C(x)|$

$$\sum_{x \in \mathcal{X}} 2^{-\ell(x)} \leq 1 \qquad (*)$$

moreover, for any $\{\ell(x)\}$ satisfying this, we can find a prefix code

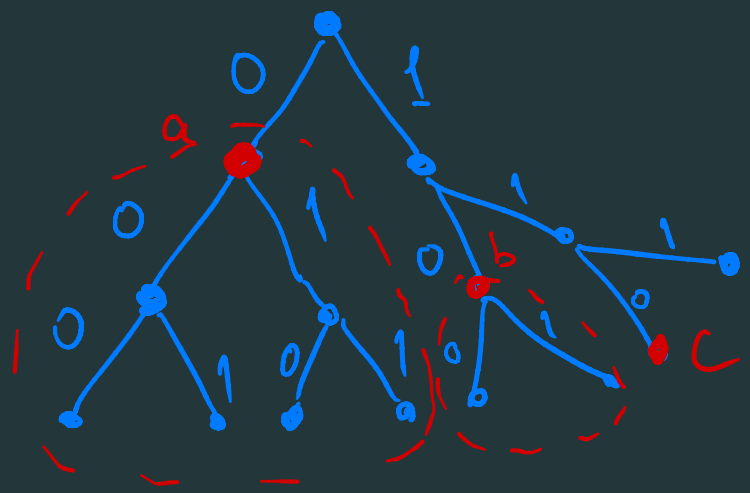
Kraft's Ineq - Prefix code iff (*)

McMillan - (*) true for any uniquely decodable code

the limits of unique decodability

Kraft's inequality: for prefix codes $\sum_{x \in \mathcal{X}} 2^{-\ell(x)} \leq 1$

Coding tree -



$C = \{\underset{a}{0}, \underset{b}{10}, \underset{c}{110}\}$

Let l_{max} be max length

- # of leaves $= 2^{l_{max}}$
- Each codeword of length l_i 'eats up' $2^{l_{max}-l_i}$ leaves

$$\Rightarrow \sum_x 2^{l_{max}-\ell(x)} \leq 2^{l_{max}}$$

Kraft's symbol-code supermarket

Kraft's inequality: for prefix codes $\sum_{x \in \mathcal{X}} 2^{-\ell(x)} \leq 1$

Cost 2^{-1} 2^{-2} 2^{-3} 2^{-4}

0	00	000	0000	The total symbol code budget
			0001	
		001	0010	
			0011	
	01	010	0100	
			0101	
		011	0110	
			0111	
1	10	100	1000	
			1001	
		101	1010	
			1011	
	11	110	1100	
			1101	
		111	1110	
			1111	

Budget = 1

Kraft's symbol-code supermarket

• Note - If C satisfies $\sum 2^{-l_i} = 1$, then 'complete code'

C_0

0	00	000	0000
			0001
		001	0010
			0011
	01	010	0100
			0101
		011	0110
			0111
1	10	100	1000
			1001
		101	1010
			1011
	11	110	1100
			1101
		111	1110
			1111

C_3

0	00	000	0000
			0001
		001	0010
			0011
	01	010	0100
			0101
		011	0110
			0111
1	10	100	1000
			1001
		101	1010
			1011
	11	110	1100
			1101
		111	1110
			1111

C_4

0	00	000	0000
			0001
		001	0010
			0011
	01	010	0100
			0101
		011	0110
			0111
1	10	100	1000
			1001
		101	1010
			1011
	11	110	1100
			1101
		111	1110
			1111

C_6

0	00	000	0000
			0001
		001	0010
			0011
	01	010	0100
			0101
		011	0110
			0111
1	10	100	1000
			1001
		101	1010
			1011
	11	110	1100
			1101
		111	1110
			1111

a	0	00	0
b	10	01	01
c	110	10	011
d	111	11	111

complete codes

optimizing expected code length

let $X \sim \{p(x)\}_{x \in \mathcal{X}}$, and consider code $C(\cdot)$, and let $\ell(x) = |C(x)|$
the expected length of C is $\mathbb{E}[L(C, X)] = \sum_x p(x)\ell(x)$

$$\max \sum_{x \in \mathcal{X}} p(x)\,\ell(x) \qquad \text{s.t.} \quad \sum_{x \in \mathcal{X}} 2^{-\ell(x)} \leq 1$$

$$\ell(x) \in \{1, 2, \dots\}$$

- Define : $q(x) = \dfrac{2^{-\ell(x)}}{\sum_{y \in \mathcal{X}} 2^{-\ell(y)}} = \dfrac{2^{-\ell(x)}}{Z}$

$$\Rightarrow \mathbb{E}[L(C,X)] = \sum_{x \in \mathcal{X}} p(x) \log_2 \left(\frac{1}{Z\,q(x)} \cdot \frac{p(x)}{p(x)} \right)$$

$$= H(X) + \sum_{x \in \mathcal{X}} p(x) \log_2 \left(\frac{p(x)}{q(x)} \right) - \log_2 Z$$

relative entropy and Gibb's inequality

relative entropy (or Kullback-Leibler (KL) divergence)

the relative entropy $D_{KL}(p||q)$ between two distributions $p(x)$ and $q(x)$ defined over alphabet $\mathcal{X}$ is

$$D_{KL}(P||Q) = \sum_{x \in \mathcal{X}} p(x) \ln \left(\frac{p(x)}{q(x)} \right)$$

$$= -\sum p(x) \ln q(x) - H(x)$$

relative entropy and Gibb's inequality

the relative entropy $D_{KL}(p||q) = \sum_{x\in\mathcal{X}} p(x) \ln\left(\frac{p(x)}{q(x)}\right) \geq 0$ for all p, q

- Jensen's Inequality
  - $g(x) = x \ln x$, $g'(x) = 1 + \ln(x)$ ↑
    $\Rightarrow g(x)$ is convex
  - $-\sum_x q(x)\left[\left(\frac{p(x)}{q(x)}\right)\ln\left(\frac{p(x)}{q(x)}\right)\right] \geq \left(\sum_x \frac{p(x)}{q(x)} \cdot q(x)\right) \ln\left(\sum \frac{p(x)q(x)}{q(x)}\right)$
    $g(p(x)/q(x))$ $= 0$

optimizing expected code length

- entropy of X: $H(X) = \sum_{i \in \mathcal{X}} p_i \log_2 \left(\frac{1}{p_i}\right)$
- Kraft-McMillan inequality: UD code $\{\ell_i\}_{\{i \in \mathcal{X}\}}$ satisfies $\sum_{i \in \mathcal{X}} 2^{-\ell_i} \leq 1$

Given codeword lengths ℓ_i, define 'implicit probability'

$$\forall\, i \in \mathcal{X}, \quad q_i \triangleq \frac{2^{-\ell_i}}{\underbrace{\sum_{j \in \mathcal{X}} 2^{-\ell_j}}_{Z}} \Rightarrow \ell_i = -\log_2 q_i - \log_2 Z$$

Z = normalizing constant (partition function)

- From Kraft, $Z \leq 1 \Rightarrow \log_2 Z \leq 0$, 0 only for complete codes
- $\mathbb{E}[L(C,X)] = \sum_i p_i \ell_i = \left(\sum_i -p_i \log q_i\right) - \log_2 Z$

optimizing expected code length

let $X \sim \{p(x)\}_{x \in \mathcal{X}}$, and consider code $C(\cdot)$, and let $\ell(x) = |C(x)|$
the expected length of C is $\mathbb{E}[L(C, X)] = \sum_x p(x)\ell(x)$

- Aim - $\min \sum_i p_i l_i \quad \text{s.t} \quad \sum_i 2^{-l_i} \leq 1$

$$\Rightarrow \min \underbrace{\sum_i p_i \log_2(1/q_i) - \log_2 z}, \quad \log_2 z \leq 0$$

$$= H(p) + \sum_i p_i \Big(\underbrace{\log_2 p_i + \log_2(1/q_i)}_{\log_2(p_i/q_i)} \Big) + \log_2\left(\frac{1}{z}\right)$$

$$\Rightarrow \mathbb{E}[L(C,X)] = H(p) + \underbrace{D_{KL}(p \| q)}_{\text{Kullback-Leibler Divergence}} + \log_2(1/z)$$

relative entropy and Gibb's inequality

relative entropy (or Kullback-Leibler (KL) divergence)

the relative entropy $D_{KL}(p||q)$ between two distributions $p(x)$ and $q(x)$ defined over alphabet $\mathcal{X}$ is

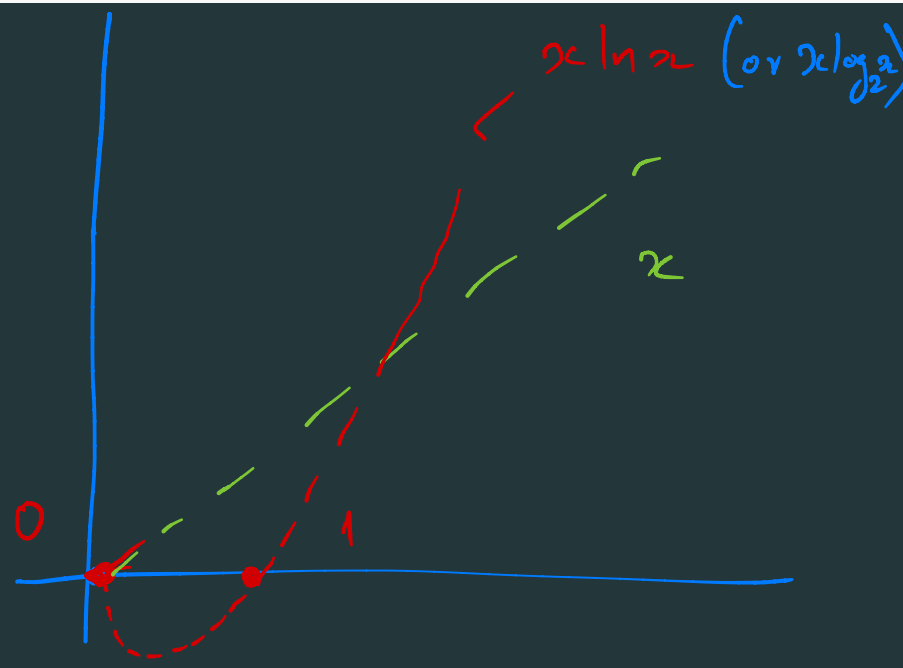
$$D_{KL}(P||Q) = \sum_{x \in \mathcal{X}} p(x) \ln \left(\frac{p(x)}{q(x)} \right)$$

Think of it as a 'distance' between P, Q

- Fact - $D_{KL}(P||P) = 0$
- Fact - $D_{KL}(P||Q) \neq D_{KL}(Q||P)$
- (Not yet) Fact - $D_{KL}(P||Q) \geq 0$

the function $\phi(x) = x \ln x$ ~~$\ln x$~~ $\log_2 x$

- $\lim_{x \to 0} \left(\frac{\ln x}{1/x} \right) = \lim_{x \to 0} \left(\frac{1/x}{-1/x^2} \right) = 0$



- $\frac{d\phi(x)}{dx} = 1 + \ln x$

$\frac{d^2\phi(x)}{dx^2} = \frac{1}{x} > 0 \; \forall x \geq 0 \Rightarrow \phi(x)$ is convex

- Jensen's Ineq $\mathbb{E}[\phi(x)] \geq \phi(\mathbb{E}[x])$

relative entropy and Gibb's inequality

the relative entropy $D_{KL}(p||q) = \sum_{x\in\mathcal{X}} p(x) \ln\left(\frac{p(x)}{q(x)}\right) \geq 0$ for all p, q

- $D_{KL}(p||q) = \sum_x q(x) \left(\frac{p(x)}{q(x)}\right) \ln\left(\frac{p(x)}{q(x)}\right)$ (ie, $q(x) > 0 \; \forall x$ s.t $p(x) > 0$)

p absolutely cont wrt q

Define $Y = p(x)/q(x)$ with prob $q(x)$ $\forall x \in \mathcal{X}$

$\Rightarrow D_{KL}(p||q) = \mathbb{E}_q[Y \ln Y]$

$\geq (\mathbb{E}_q Y) \ln(\mathbb{E}_q Y) \geq 0$

$\mathbb{E}_q(Y) = \sum_x q(x) \cdot \left(\frac{p(x)}{q(x)}\right) = 1$

optimizing expected code length

$$\mathbb{E}[L(C,X)] = H(P) + D_{KL}(P\|Q) + \log_2 \tfrac{1}{Z}$$

(where $Z = \sum_i 2^{-\ell_i}$, $D_{KL}(P\|Q) = \sum_i p_i \log\left(\frac{p_i}{q_i}\right)$)

$$\geq H(P) \quad \text{(Via Kraft + Gibbs)}$$

Moreover, $\mathbb{E}[L(C,X)] = 0$ if $Z = 1$ (ie complete code)

and $p_i = q_i$

$$\Rightarrow \ell_i = \log_2\left(\frac{1}{p_i}\right) = h(a_i)$$

optimizing expected code length

In practice - choose $\ell_i = \lceil \log_2 \frac{1}{p_i} \rceil$

- (Check) $\sum_{i \in X} 2^{-\ell_i} \leq 1$

- $\mathbb{E}[L(C,X)] = \sum_i p_i \overbrace{\lceil \log_2 \frac{1}{p_i} \rceil}^{\leq \left(\log_2 \frac{1}{p_i}\right) + 1}$

$\leq H(p) + 1$

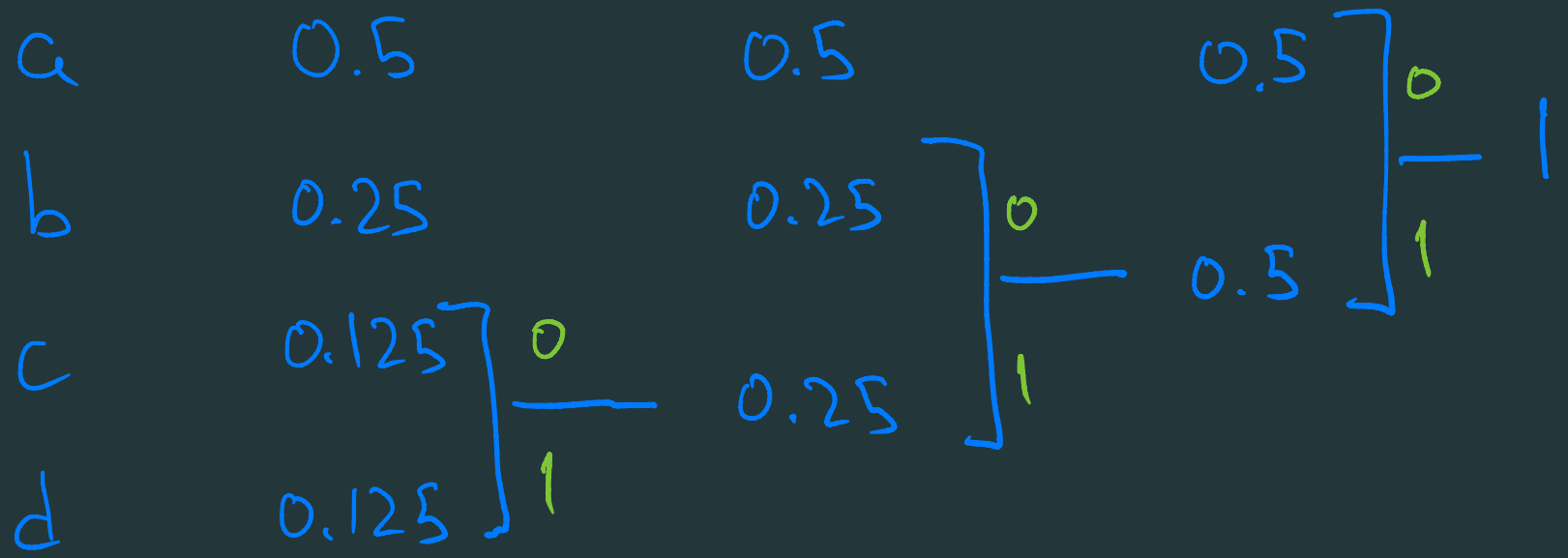
aside: cross entropy

the cross entropy of p given q: $H_p(q) = \sum_{x \in \mathcal{X}} p(x) \ln\left(\frac{1}{q(x)}\right)$
– avg length of message from if 'p mis-estimated as q'

- $D_{KL}(P \| Q) = \sum P_i \log \frac{P_i}{q_i} \geq H(P) = \sum P_i \log \frac{1}{P_i}$
- $H_p(Q) = D_{KL}(P \| Q) + H(P)$ (?)

Huffman code

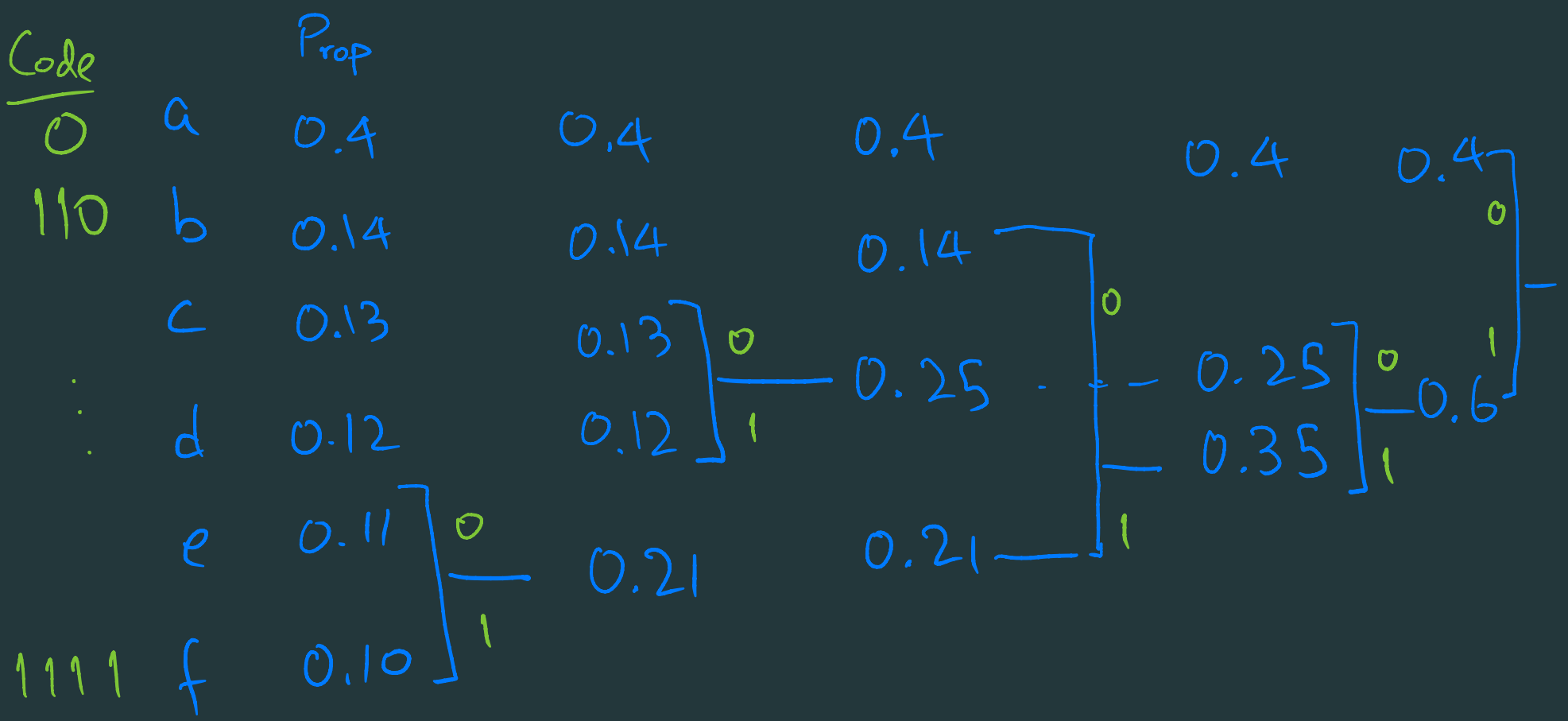
consider $X \sim \{a, b, c, d, \text{[illegible]}\}$ with prob $\{0.5, 0.25, 0.125, 0.125\}$



$$\Rightarrow \{0, 10, 110, 111\}$$

Huffman code

consider $X \sim \{a, b, c, d, e, f\}$ with prob $\{0.4, 0.14, 0.13, 0.12, 0.11, 0.10\}$



aside: information content in a perfect code

let C be a perfect code for X, and given database $X_1X_2\ldots X_n$, suppose we pick one bit at random from the encoded sequence $C(X_1)C(X_2)\ldots C(X_n)$. what is the probability this bit is a 1?

P	alphabet	code	'# of 1s'
1/2	a	0	0
1/4	b	10	1/2
1/8	c	110	2/3
1/8	d	111	1

$$\mathbb{P}[\text{bit}=1] = 0 + 1/8 + 2/24 + 1/8$$

$$= 1/3$$

$$= \sum p_i f_i \quad \times$$

$$\mathbb{P}[\text{bit}=1] = \frac{\sum p_i f_i \cdot \ell_i}{\sum p_i \ell_i} = 0.5$$