Problem sheet 3, Information Theory, HT 2022

Designed for the third tutorial class

Question 1 For a random variable X with state space $X = \{x_1, \cdots, x_7\}$ and distribution $p_i = \mathbb{P}(X = x_i)$ given by

p_1	p_2	p_3	p_4	p_5	p_6	p_7
0.49	0.26	0.12	0.04	0.04	0.03	0.02

(a) Find a binary Huffman code for X and its expected length.

(b) Find a ternary Huffman code for X and its expected length.

Answer 1 For the binary Huffman,

Table 1: Binary Huffman

step 1	$p_1 = 0.49$	$p_2 = 0.26$	$p_3 = 0.12$	$p_4 = 0.04$	$p_5 = 0.04$	$p_6 = 0.03$	$p_7 = 0.02$
step 2	$p_1 = 0.49$	$p_2 = 0.26$	$p_3 = 0.12$	$p_4 = 0.04$	$p_5 = 0.04$	$p_{67} = 0.05$	
step 3	$p_1 = 0.49$	$p_2 = 0.26$	$p_3 = 0.12$	$p_{45} = 0.08$		$p_{67} = 0.05$	
step 4	$p_1 = 0.49$	$p_2 = 0.26$	$p_3 = 0.12$		$p_{4567} = 0.13$		
step 5	$p_1 = 0.49$	$p_2 = 0.26$		$p_{34567} = 0.25$			
step 6	$p_1 = 0.49$		$p_{234567} = 0.516$				
step 7		$p_{1234567} = 1$					

So

$$c(1) = 0, c(2) = 10, c_3 = 110, c_4 = 11100, c_5 = 11101, c_6 = 11110, c_7 = 11111;$$

and its expected length is $0.49 * 1 + 0.26 * 2 + 0.12 * 3 + 0.04 * 5 + 0.04 * 5 + 0.03 * 5 + 0.02 * 5 = 2.02$

For ternary Huffman, see the calculation in table 2.

So

$$c(1) = 0, c(2) = 1, c_3 = 20, c_4 = 21, c_5 = 220, c_6 = 221, c_7 = 222;$$

and its expected length is $(0.49 + 0.26) * 1 + (0.12 + 0.04) * 2 + (0.04 + 0.03 + 0.02)0.12 * 3 = 1.34$

Table 2: Ternary Huffman

step 1	$p_1 = 0.49$	$p_2 = 0.26$	$p_3 = 0.12$	$p_4 = 0.04$	$p_5 = 0.04$	$p_6 = 0.03$	$p_7 = 0.02$
step 2	$p_1 = 0.49$	$p_2 = 0.26$	$p_3 = 0.12$	$p_4 = 0.04$		$p_{567} = 0.09$	
step 3	$p_1 = 0.49$	$p_2 = 0.26$		$p_{34567} = 0.25$			
step 4			$p_{123456} = 1$				

Question 2 (a) Prove that the Shannon's code is a prefix code and calculate bounds on its expected length. Give an example to demonstrate that it is not an optimal code.

(b) Prove that the Elias code is a prefix code and calculate bounds on its expected length. Is it an optimal code?
Hint: Suppose $\mathcal{Y} = \{0, 1, \cdots, d\}$. For any $i = 1, \cdots, |\mathcal{X}|$, suppose $c(x_i) = a_1 \cdots a_k$ with $k = |c(x)|$. Denote $v_i = \sum_{j=1}^{|c(x_i)|} a_j d^{-j}$, $r(i) = \sum_{j=1}^{i-1} p_j + p_i/2$ and $\hat{r}(i) = r(i) + p_i/2$. Try to show that the interval $[v_i, v_i + d^{-|c(x_i)|})$ is contained in the interval $[\hat{r}_{i-1}, \hat{r}_i)$. Hence the intervals $[v_i, v_i + d^{-|c(x_i)|})$ are disjoint to each other.

Answer 2 (a) For the Shannon's code, its length $l_x = \lceil -\log(p_X(x)) \rceil$, which satisfies $\sum_x d^{-l_x} \leq \sum_x d^{\log(p_X(x))} = \sum_x p_X(x) = 1$. So l_x satisfies the Kraft-McMillan's inequality. Furthermore, the Shannon's code is exactly the one constructed in the proof of Theorem 3.5 in the elcture notes, so it is a prefix code.

A counter example for the optimality of Shannon's code is as follows: $\mathcal{X} = \{\mathcal{A}, \mathcal{B}\}$, $p_X(B) = 2^{-4}, p_X(A) = 1 - 2^{-4}$, then

$$p_1 = P_X(A) = 1 - 2^{-4}, p_2 = 2^{-4},$$

so

$$r_1 = 0, r_2 = p_1, l_1 = 1, l_2 = 4, c(A) = 0, c(B) = 1111.$$

This is obviously not optimal since $c(A) = 0, c(B) = 1$ is strictly better.

(b) Without loss of generality, suppose $\mathcal{X} = \{1, 2, \cdots, m\}$ and $\mathcal{Y} = \{0, 1, \cdots, d\}$. The distribution of X is $p_i = \mathbb{P}(X = i)$.

Denote $r_i = \sum_{j=1}^{i-1} p_j + \frac{p_i}{2}$ and $\hat{r}_i = \sum_{j=1}^{i} p_j = r_i + \frac{p_i}{2}$. For any fixed $i \in \mathcal{X}$, denote the d-ary expansion of r_i as $0.a_1 a_2 \cdots$ and $l_i = \lceil -\log_d(p_i) \rceil + 1$, Then the Elias code is $c(i) = a_1, \cdots a_{l_i}$.

Denote v_i as the value of d-ary expansion $0.a_1 \cdots a_{l_i}$, i.e., $v_i = \sum_{j=1}^{l_i} a_j d^{-j}$, then

$$v_i \leq r_i.$$

Together with $d^{-l_i} \leq p_i d^{-1}$ we know

$$v_i + d^{-l_i} \leq r_i + \frac{p_i}{d} \leq r_i + \frac{p_i}{2} = \hat{r}_i.$$

On the other hand, $0 \leq r_i - v_i < d^{-l_i}$, so $v_i > r_i - d^{-l_i} \geq r_i - p_i d^{-1} > r_i - p_i/2 = \hat{r}_{i-1}$.

So we have $[v_i, v_i + d^{-l_i}) \subseteq [\hat{r}_{i-1}, \hat{r}_i)$, which implies $[v_i, v_i + d^{-l_i})$ are disjoint.

If c is not a prefix code, then $\exists\, i \neq j$ such that $c(i) = c(j)y$ for some $y \in \mathcal{Y}^*$, hence $v_i = v_j + d^{-l_j} z$ with the d-ary expansion of z being $0.y$. So $v_i \in [v_j, v_j + d^{-l_j})$. But this is impossible because $[v_i, v_i + d^{-l_i})$ are disjoint.

Question 3 Prove the following weaker version of the Kraft-McMillan theorem (called Krafts theorem) using rooted trees

(a) Let $c : \mathcal{X} \mapsto \{0, \cdots, d-1\}^*$ be a prefix code. Consider its code-tree and argue that $\sum_{x \in \mathcal{X}} d^{-|c(x)|} \leq 1$. [Note that the assumption that c is a prefix code is crucial here, otherwise the code-tree cannot be defined to begin with. In the Kraft-McMillan theorem from the lecture we only require c to be uniquely decodable].

(b) Assume that $\sum_{x \in \mathcal{X}} d^{-l_x} \leq 1$ with $l_x \in \mathbb{N}$. Show that there exists a prefix code c with codeword lengths $|c(x)| = l_x$ for $x \in \mathcal{X}$ by constructing a rooted tree.

Answer 3 A prefix code is equivalent to a rooted tree, where each codeword corresponds to a path from a leave to the root.

(a) We call a d-ary tree being semi-complete if the degree of every non-leave vertex has d direct descendants. In a semi-complete d-ary tree for any leave x, denote $h(x)$ as the height from the root to the tree with $h(root) = 0$. It is easy to check that $\sum_{\text{every leave } x} d^{-h(x)} = 1$.

For the code-tree of a prefix code, it can be expanded to a semi-complete tree by adding some leaves to a non-leave vertex. Hence $\sum_{\text{every leave } x} d^{-h(x)} \leq 1$.

(b) We call a d-ary tree being complete with height h if it is semi-complete, the distance from each leave to the root is h.

Given l_x satisfies the condition, denote $h = \max_x l_x$, then we can construct a d-ary complete tree with maximal height H.

Suppose $l_1 \leq l_2 \leq \cdots \leq l_m$. We mark nodes and cut branches of a complete tree as follows:

(1) Take $i = 1$.

(2) Find the first non-marked node on the left of the tree with height l_i, cut off its descendant vertices, and mark all ancestral vertices (including itself) and their edges up to the root.

(3) Set $i = i + 1$ and repeat (2) until $i = m + 1$.

By the assumption $\sum_{i=1}^{m} d^{-l_i} \leq 1$, we know we can run this construction for all $k \leq m$ (otherwise, if we cannot find a node with height l_k at some $k \leq m$, then it must happen that $\sum_{i=1}^{k} d^{-l_x} > 1$). All marked vertices and edges and the i^{th} leave in the algorithm corresponds to the codeword i.

Question 4 Give yet another proof for $\sum_{x\in\mathcal{X}} d^{-|c(x)|} \leq 1$ if c is a prefix code by using the "probabilistic method": randomly generate elements of $\{0,\cdots,d-1\}^*$ by sampling i.i.d. from $\{0,\cdots,d-1\}$ and consider the probability of writing a codeword of c.

Answer 4 Sample independent uniform variables on $\{0,\cdots,d-1\}$, append them inductively to the right, and stop if we obtain a codeword of c. By definition, and since c is a prefix code, the probability of writing the word $c(x)$ is exactly $d^{-|c(x)|}$. Thus the probability for this process to stop is equal to $\sum_{x\in\mathcal{X}} d^{-|c(x)|}$, implying this quantity is at most 1.

Question 5 Let X be uniformly distributed over a finite set $\mathcal{X}$ with $|\mathcal{X}| = 2^n$ for some $n \in \mathbb{N}$. Given a sequence $A_1, A_2, \cdots$ of subsets of $\mathcal{X}$ we ask a sequence of questions of the form $X \in A_1, X \in A_2$, etc.

(a) We can choose the sequence of subsets. How many such questions do we need to determine the value of X? What is the most efficient way to do so?
`[Note: If we regard all questions as a mapping from` $\mathcal{X}$ `to` $\{Yes, No\}^*$`, we can even think about how to design the sequence of subsets to minimise the expected number of questions to ask to get the value of a random variable` X `with any given distribution.]`

(b) We now randomly (i.i.d. and uniform) draw a sequence of sets $A_1, A_2, \cdots$ from the set of all subset of $\mathcal{X}$. Fix $x, y \in \mathcal{X}$. Conditional on $\{X = x\}$:

i. What is the probability that x and y are indistinguishable after the first k random questions?

ii. What is the expected number of elements in $\mathcal{X}\backslash\{x\}$ that are indistinguishable from x after the first k questions?

Answer 5 (a) Huffman codes are of length n, hence we can identify X in n deterministic(!) questions.

(b) Notice that a uniform random subset of $\mathcal{X}$ contains each $x \in \mathcal{X}$ independently with probability $1/2$.

i. The probability that a random subset A distinguishes x and y is $1/2$. Since the A_i are independent, the probability is 2^{-k}.

ii. For all $y \neq x$, let B_y be 1 if the questions $A_1, \cdots, A_k$ do not distinguish x and$y$, and 0 if they do. Then the (B_y) are all Bernouilli random variables with parameter 2^{-k}, and there are 2^{n-1} of them. The wanted expectation is $\mathbb{E}\left[\sum_{y\neq x} B_y\right] = (2^n - 1)2^{-k}$.

Question 6 Let $|\mathcal{X}| = 100$ and p the uniform distribution on $\mathcal{X}$. How many codewords are there of length $l = 1, 2, \cdots$ in an Huffman binary code?

Answer 6 By Huffman procedure, we can see that there are 28 codewords of length 6 and 72 of length 7.

Another way to get this numbers is as follows:

Consider the optimisation of l_x for optimal code

$$\min \sum_{i=1}^{100} p_i l_i \quad \text{subject to } \sum_{i=1}^{100} 2^{-l_i}$$

The optimal l_i should be integers close to $-\log(p_i)$, which is 6 or 7 in this question.

To prove this, denote $\Gamma = \{u = (u_1, \cdots u_{100}) : \sum 2^{-u_i} \leq 1\}$ being the set of feasible solutions (without integer constraint), and $J(u) = \sum u_i$ being the objective function.

Denote $u^* = \log(100) * (1, 1, 1, \cdots 1)$, $A = \{6, 7\}^{100} \cap \Gamma$, and $\bar{A}$ be convex hull of A, which is contained in Γ. Then for any feasible solution out of $\bar{A}$, the segment between u and u^* must intersect with $\bar{A}$, hence intersect with the surface of $\bar{A}$. So, there exists a $\lambda \in (0, 1)$ such that $u^\lambda = \lambda u + (1-\lambda)u^*$ is on the surface of $\bar{A}$, and $J(u^\lambda) = \lambda J(u) + (1-\lambda)J(u^*)$. Since $J(u^*) < J(u^*)$, so $J(u^\lambda) < J(u)$. Furthermore, $u^*\lambda$ is on the surface of $\bar{A}$, so there exists a $\hat{u} \in A$ such that $J(\hat{u}) \leq J(u^\lambda)$, which implies u cannot be optimal.

Since $p_i = 1/100$, $\log(p_i) \in (6, 7)$, so l_x can only be 6 or 7. Suppose there are k 7's and $100 - k$ 6's, then $k2^{-7} + (100 - k)2^{-6} = 100 * 2^{-6} - k * 2^{-7} \leq 1$ and we want k to be as big as possible. Hence $k = \text{floor}(2^7(100 * 2^{-6} - 1)) = \text{floor}(200 - 128) = 72$.

Question 7 (Optional) Let X be a Bernoulli random variable with $\mathbb{P}(X = 0) = 0.995, \mathbb{P}(X = 1) = 0.005$ and consider a sequence $X_1, \cdots, X_{100}$ consisting of i.i.d. copies of X. We study a block code of the form $c : \{0, 1\}^{100} \mapsto \{0, 1\}^m$ for a fixed $m \in \mathbb{N}$.

(a) What is the minimal m such that there exists c such that its restriction to sequences $\{0, 1\}^{100}$ that contain three or fewer 1s is injective?

(b) What is the probability of observing a sequence that contains four or more 1s? Compare the bound given by the Chebyshev inequality with the actual probabiltiy of this event.

Answer 7 (a) n The number of binary sequences with 3 or fewer ones is

$$\begin{pmatrix} 100 \\ 0 \end{pmatrix} + \begin{pmatrix} 100 \\ 1 \end{pmatrix} + \begin{pmatrix} 100 \\ 2 \end{pmatrix} + \begin{pmatrix} 100 \\ 3 \end{pmatrix} = 166751,$$

so the required minimal codeword length is

$$\lceil \log_2(166751 \rceil = 18.$$

(b) The probability of having at most 3 ones is

$$\sum_{i=0}^{3} \begin{pmatrix} 100 \\ i \end{pmatrix} 0.05^i (0.995)^{100-i} \approx 0.99833$$

and the wanted probability is approximately $1 - 0.99833 = 0.00167$.

It is easy to check $\mathbb{E}[X] = 0.005$ and $\mathrm{Var}(X) = 0.995 * 0.005^2 + 0.005 * 0.995^2 = 0.005 * 0.995 \approx 0.005$.

Denote $\bar{X} = \sum_{i=1}^{100}(X_i - \mathbb{E}[X_i])$, then $\mathbb{E}[\bar{X}] = 0$, $\mathbb{E}[\bar{X}^2] = \mathrm{Var}(\bar{X}) = 100\mathrm{Var}(X) = 0.5$. Recall that Chebyshevs inequality states that

$$\mathbb{P}\left(|\bar{X}| \geq \varepsilon\right) \leq \frac{\mathbb{E}[\bar{X}^2]}{\varepsilon^2} = \frac{0.5}{\varepsilon^2}.$$

Now we want to estimate the probability for

$$\sum_{i=1}^{100} X_i \geq 4 \Leftrightarrow \bar{X} \geq 4 - 0.5 = 3.5.$$

Hence we take $\varepsilon = 3.5$, then

$$\mathbb{P}(\sum_{i=1}^{100} X_i \geq 4) = \mathbb{P}(\bar{X} \geq 3.5) < \mathbb{P}(|\bar{X}| \geq 3.5) \leq \frac{0.5}{3.5^2} \approx 0.0406.$$

In fact, if we use the central limit theory, we know $\frac{\bar{X}}{\sqrt{100\mathrm{Var}(X)}} = \bar{X}\sqrt{2}$ approximately follows the standard normal, then $\mathbb{P}(\bar{X} \geq 3.5) = \mathbb{P}(\bar{X}\sqrt{2} \geq 7/\sqrt{2} = 1 - \Phi(7/\sqrt{2}) \approx 3.7 * 10^{-7}$.

Question 8 (Optional) Let X be a $\mathcal{X} = \{1, 2, 3, 4\}$-valued random variable with pmf p and binary code c as in the Table 1.

x=	1	2	3	4
p=	0.5	0.25	0.125	0.125
c=	0	10	110	111

Table 3: Data for Question 8

.

For $n \in \mathbb{N}$, we generate a sequence in $\mathcal{X}^n$ by sampling i.i.d. from the distribution p. We then pick one bit uniformly at random from the binary encoded sequence. What is the asymptotic (as $n \to +\infty$) probability that this bit equals 1?

Answer 8 Let $X_1, \cdots, X_n$ be i.i.d. copies of X. For each i, let Y_i be the number of ones in $c(X_i)$ and Z_i the number of bits in $c(X_i)$.

For a fixed n, the wanted probability is, equal to

$$\frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n Z_i} = \frac{\sum_{i=1}^n Y_i/n}{\sum_{i=1}^n Z_i/n}.$$

When $n \to$, by the SLLN, we have the a.s. convergences

$$\sum_{i=1}^n Y_i/n \to \mathbb{E}[Y_1] = 7/8, \quad \sum_{i=1}^n Z_i/n \to \mathbb{E}[Z_1] = 7/4.$$

Hence the asymptotic probability is $1/2$.