Introduction to Information and Coding Theory

Lecture 4

Lecturer: Haim Permuter *Scribe: Amitai Koretz and Tamir Harush*

I. Optimality of Huffman Codes

Lemma 1 (Canonical) For any probability distribution $[p_1, p_2, ..., p_m]$, there exists an optimal prefix code which satisfies the following properties:

1) if $p_i \geq p_j$, then $l_i \leq l_j$.
2) The two largest codewords are of the same length.
3) The two largest codewords differ only in the last bit.

A code which fulfils the above properties is termed *canonical*.

Proof:

1) Let us assume that the opposite is true. Namely, $l_i \geq l_j$. The contribution to the expected length of these two codewords is $(l_i p_i + l_j p_j)$. Now, let us exchange between the codewords such that l_i will be associated with symbol j and l_j with symbol i. The contribution to the expected length of these two codewords is $(l_i p_j + l_j p_i)$. Now consider the difference

$$\begin{aligned}(l_i p_i + l_j p_j) - (l_i p_j + l_j p_i) &= l_i(p_i - p_j) + l_j(p_j - p_i) && (1)\\ &= (l_i - l_j)(p_i - p_j) \geq 0 && (2)\end{aligned}$$

since by assumption $(l_i - l_j) \geq 0$ and $(p_i - p_j) \geq 0$. Hence, if $l_i \geq l_j$ the code is not optimal and we obtained a contradiction.

2) Assume the two largest codewords are not of the same length. Without changing the expected codeword length, the larger of the two codewords may be truncated of its extra bits so that it the same length as its sibling. The two must still be distinct since the assumption is that the code is prefix code.
3) According to the previous item, the two largest codewords are of the same length. Without loss of generality, that the last codeword may be changed so that they differ

only in the last bit. Using the previous property we can conclude the codewords of the two symbols with the lowest probability have the same codeword length, and the differ only in the last bit.

■

Theorem 1 (Optimality of Huffman code) Huffman codes are optimal, i.e., having the lowest average length-code..

Proof: Let $C^*_m(p_1, p_2, \ldots, p_{m-1}, p_m)$ be an optimal canonical code such that $p_1 \geq p_2 \geq \cdots \geq p_m$ and denote the ith codeword by $C^*_{m,i}$. Now define a new code C_{m-1} using the following definition:

$$C_{m-1} = \begin{cases} C_{m-1,i} = C^*_{m,i} & \text{, for } i \leq m-2 \\ C_{m-1,m-1} = \text{merge}(C^*_{m,m-1}, C^*_{m,m}) & \text{, for } i = m-1 \end{cases} \tag{3}$$

where the merge$(\cdot,\cdot)$ function truncates the bit which differentiates the two largest codewords. Note that eq. (3) defines a construction of code C_{m-1} from C_m and also C_m from C_{m-1} which exactly as one step in the Huffman code. Note that C_{m-1} is a prefix code.

In order to prove that Huffman code procedure obtains an optimal prefix code (namely a prefix code with smallest expected length) we need to show that if C_m is an optimal canonical prefix code then also C_{m-1} is an optimal prefix code. If it is, we can apply the procedure above (which is the procedure we do in the Huffman code in one step) again and again till we get a code with only two codewords, and trivially this prefix code is $\{0, 1\}$.

The expected codeword length of C_{m-1} is calculated as:

$$\begin{aligned} L(C_{m-1}) &= \sum_{i=1}^{m-2} p_i l_i + p_{m-1}(l_{m-1} - 1) + p_m(l_m - 1) && (4) \\ &= L(C^*_m) - p_{m-1} - p_m && (5) \end{aligned}$$

where l_i is the length of the ith codeword. Therefore, we have that:

$$L(C^*_m) = L(C_{m-1}) + p_m + p_{m-1}. \tag{6}$$

It follows that C_{m-1} is an optimal prefix code, since were that not the case, we can improve the expected codeword length of C_m^* by replacing C_{m-1} with another canonical prefix code that is an optimal code. This contradicts the optimality of C_m^*. Therefore, C_{m-1} is optimal, and we may assume without loss of generality that it is canonical as well.

■

II. Lower bound of the expected length of Decodable Codes via Kraft's inequality

We have previously seen that Kraft's inequality, i.e.

$$\sum_i 2^{-l_i} \leq 1 \tag{7}$$

holds for all prefix codes and we saw that a direct consequence of the inequality is that $\mathbb{E}[l(X)] \geq H(X)$. Now we will show that the above holds for any uniquely decodable code.

Recall the definition of *uniquely decodable codewords*: We say that a code is uniquely *decodable*, if any extension of codewords is non-singular, where an extension of codewords is a concatenation $f(x_1)f(x_2)f(x_3)f(x_4)\dots$ without any spaces or commas. Recall that a code is *non-singular* if for any $x_1 \neq x_2 \ \Rightarrow f(x_1) \neq f(x_2)$.

Theorem 2 (Kraft inequality for uniquely decodable code) Kraft's inequality is satisfied for any uniquely decodable code.

Proof:

Assume a uniquely decodable code and an alphabet $\mathcal{X}$. Then

$$\begin{aligned}
\left(\sum_{x\in\mathcal{X}} 2^{-l(x)}\right)^k &\overset{(a)}{=} \sum_{x_1\in\mathcal{X}}\sum_{x_2\in\mathcal{X}}\cdots\sum_{x_k\in\mathcal{X}} 2^{-l(x_1)}2^{-l(x_2)}\dots 2^{-l(x_k)} \\
&\overset{(b)}{=} \sum_{x^k\in\mathcal{X}^k} 2^{-l(x^k)} \\
&\overset{(c)}{=} \sum_{m=1}^{k\cdot l_{max}} a(m)2^{-m},
\end{aligned} \tag{8}$$

where $a(m)$ is the number of sequences $x^k \in \mathcal{X}^k$ such that $l(x^k) = m$ and l_{max} is the maximum codeword length. Step (a) follows the fact that $\left(\sum_{x \in \mathcal{X}} x\right)\left(\sum_{y \in \mathcal{Y}} y\right) = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} xy$, Step (b) from the fact that the code encode each input signal x_i separately and therefore $l(x^k) = l(x_1) + l(x_2) + ... + l(x_k)$ and Step (c) from changing the summation from $x^k \in \mathcal{X}^k$ to the total length of code that is bounded between 1 (or even $1 \cdot k$) to $k \cdot l_{max}$.

Since the code is uniquely decodable, it must hold that $a(m) \leq 2^m$ (otherwise there are at least two symbols with the same codeword). Therefore for any k:

$$\left(\sum_{x \in \mathcal{X}} 2^{-l(x)}\right)^k = \sum_{m=1}^{k \cdot l_{max}} a(m) 2^{-m} \leq \sum_{m=1}^{k \cdot l_{max}} 2^m 2^{-m} = k \cdot l_{max}. \tag{9}$$

Taking the kth root on both sides of the inequality, we have:

$$\sum_{x \in \mathcal{X}} 2^{-l(x)} \leq (k \cdot l_{max})^{\frac{1}{k}}. \tag{10}$$

Since the above holds for any value of k, we let $k \to \infty$ so that $(k \cdot l_{max})^{\frac{1}{k}} \xrightarrow[k \to \infty]{} 1$ [1]. And therefore,

$$\sum_{x \in \mathcal{X}} 2^{-l(x)} \leq 1. \tag{11}$$

■

Corollary 1 For any uniquely decodable codes

$$\mathbb{E}[l(X)] \geq H(X) \tag{12}$$

Proof: Inequality (12) follows from Kraft inequality as shown previously in lecture 3. ■

III. Markov Chains

Let X, Y, Z be random variables. We denote $X \to Y \to Z$ if

$$P(x, y, z) = P(x)P(y|x)P(z|y), \tag{13}$$

[1] $\lim_{k \to \infty} (k \cdot l_{max})^{\frac{1}{k}} = \lim_{k \to \infty} e^{\frac{\ln(k \cdot l_{max})}{k}} \xrightarrow[k \to \infty]{} 1$ since k tends to infinity faster than $\ln k$.

or equivalently

$$P(z|x,y) = P(z|y). \tag{14}$$

Lemma 2 If $X \to Y \to Z$, then $X \leftarrow Y \leftarrow Z$, so that the notation $X - Y - Z$ may be substituted.

Proof: $P(z|x,y) = P(z|y) \Rightarrow P(x|y,z) = P(x|y)$, since

$$P(x|y,z) = \frac{P(x,y,z)}{P(y,z)} = \frac{P(x)P(y|x)P(z|y)}{P(y)P(z|y)} = \frac{P(x,y)}{P(y)} = P(x|y). \tag{15}$$

■

Lemma 3 (Data Processing Inequality)

$$X - Y - Z \Rightarrow I(X;Y) \geq I(X;Z). \tag{16}$$

Proof:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &\overset{(a)}{=} H(X) - H(X|Y,Z) \\ &= I(X;Y,Z) \\ &\overset{(b)}{=} I(X;Z) + I(X;Y|Z), \end{aligned} \tag{17}$$

where (a) follows from the Markov property defined by (14), and (b) follows from chain-rule.

■

Proof: (*Alternative proof for Lemma* (3))

$$I(X;Y,Z) = H(X) - H(X|Y,Z) \geq H(X) - H(X|Z) = I(X;Z) \tag{18}$$

since conditioning reduces entropy. ■

Corollary 2 Mutual information cannot be increased by a deterministic transformation.

Proof: Let $Z = g(Y)$ such that Z is a deterministic function of Y. Therefore $X - Y - g(Y)$, and according to the Data Processing inequality

$$I(X;Y) \geq I(X;g(Y)). \tag{19}$$

■

Theorem 3 (Law of Large Numbers) Let X_i be a collection of i.i.d. random variables. It holds that

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{P} \mu \tag{20}$$

where $\mu = \mathbb{E}[X]$, i.e.

$$\forall\epsilon > 0 \quad \lim_{n\to\infty} Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \epsilon\right) = 0. \tag{21}$$

Theorem 4 (Markov Inequality) If X is a non-negative random variable and $a > 0$, then

$$\Pr(X \geq a) \leq \frac{\mathbb{E}[X]}{a}. \tag{22}$$

Proof: Let

$$Y = \begin{cases} 0 & , X \leq a \\ a & , X > a. \end{cases} \tag{23}$$

Therefore $Y \leq X$ and $\mathbb{E}[Y] \leq \mathbb{E}[X]$. Since $\mathbb{E}[Y] = a \cdot P(X \geq a)$, we have that

$$a \cdot \Pr(X \geq a) \leq \mathbb{E}[X]. \tag{24}$$

■

IV. Challenge

Exercise 1 *Finding an optimal prefix code for the infinite case:*. Consider the set $\{p_i\}_{i=1}^{\infty}$, such that for any $i \in N : p_i > 0, p_i \geq p_{i+1}, \sum_{i=1}^{\infty} p_i = 1$. Find an optimal code, i.e. a prefix code such that $E[L]$ is minimal (L is the length of the code word assigned to each value of X, so L is equal to the length of the code word assigned to x_i with probability p_i). **If a student solve it (including a rigours proof of optimality) he/she gets automatically 100!**

V. AEP - Asymptotic Equipartition

We now consider randomly generated sequences of length n taken from an alphabet $\mathcal{X}$.

Definition 1 (Typical Set) We define the *typical set* $A_\epsilon^{(n)} \subseteq \mathcal{X}^n$ as the set of sequences $x^n \in \mathcal{X}^n$ such that $x^n \in A_\epsilon^{(n)}$ iff

$$H(X) - \epsilon \leq -\frac{1}{n} \log P(x^n) \leq H(X) + \epsilon. \tag{25}$$

Any sequence x^n that satisfies the above is termed a *typical sequence*.

It will be shown that when n is very large, the probability that a randomly generated i.i.d. sequence will belong to $A_\epsilon^{(n)}$ is close to one. Moreover, all sequences that are members of $A_\epsilon^{(n)}$ will have equal probability of being generated, and there are approximately $2^{nH(X)}$ such sequences in $|\mathcal{X}|^n$.

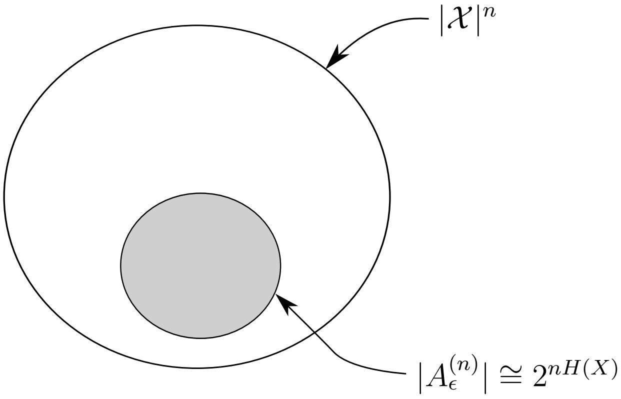

Fig. 1. Illustration of the typical set for all random i.i.d sequences of length n.

Theorem 5 Let X_i be a sequence of i.i.d. random variables. Then

$$-\frac{1}{n} \log P(X^n) \xrightarrow[n\to\infty]{P} H(X). \tag{26}$$

Proof:

$$-\frac{1}{n} \log P(X^n) \;=\; -\frac{1}{n} \log \prod_{i=1}^{n} P(X_i) = -\frac{1}{n} \sum_{i=1}^{n} \log P(X_i). \tag{27}$$

According to the Law of Large Numbers (Theorem 3), the rightmost expression converges in probability, i.e. $-\frac{1}{n}\sum_{i=1}^{n} \log P(X_i) \xrightarrow[n\to\infty]{P} H(X)$. ■