Information Theory
Lecture notes, Fall 2022

Second Lecture

September 13, 2022

Variable length source coding (cont.)

Question: Why do we care about variable length and not simply use $|\mathcal{X}|$ codewords of length $\lceil \log_s |\mathcal{X}| \rceil$ each?
Answer: Average length may be better, see this example. Let the probabilities of emitting the symbols be $p(x^{(1)}) = 1/2, p(x^{(2)}) = 1/4, p(x^{(3)}) = 1/8, p(x^{(4)})) = 1/8$. The code $f(x^{(1)}) = 0, f(x^{(2)}) = 10, f(x^{(3)})) = 110, f(x^{(4)}) = 111$ has average length $1 \cdot 1/2 + 2 \cdot 1/4 + 3 \cdot 1/8 + 3 \cdot 1/8 = 1.75 < 2 = \log_2 4$.

Aim: to minimize the *expected codelength:* $\mathbb{E}\,|f(X)| = \sum_{i=1}^{r} p_i \cdot \left| f(x^{(i)}) \right|$
However, of course we cannot assign short codewords to all source symbols and still have a uniquely decodable code.

Next we state and prove two basic theorems that belong together: they sort of complement each other. (One could certainly look at them as the two parts of a single theorem.)

Kraft-McMillan inequality

Theorem 1 *(McMillan): If $C = (f(x^{(1)}), \ldots, f(x^{(r)})$ is a UD code over an s-ary alphabet, then*

$$\sum_{i=1}^{r} s^{-|f(x^{(i)})|} \leq 1.$$

Theorem 2 *(Kraft): If the positive integers $l_1, \ldots, l_r$ satisfy*

$$\sum_{i=1}^{r} s^{-l_i} \leq 1.$$

then there exists an s-ary prefix code with codeword lengths $l_1, \ldots, l_r$.

Corollary 3 *For any uniquely decodable code, there exists a prefix code with the same codeword length.*

That means that the class of uniquely decodable codes does not offer any further choices for the set of codeword lengths than the class of prefix codes. I.e. it is enough to search for the shortest code among the prefix codes.

Proof of McMillan's theorem. Consider

$$\left(\sum_{i=1}^{r} s^{-|f(x^{(i)})|} \right)^{k} = \sum_{\mathbf{v} \in C^k} s^{-|\mathbf{v}|} = \sum_{l=1}^{k \cdot l_{\max}} A_l s^{-l},$$

where A_l is the number of ways we can have an l length string of code symbols that are concatenations of k codewords from our code, and $l_{\max}$ is the length of the longest codeword $f(x^{(i)})$. Since the code is UD, we cannot have more than s^l different source strings resulting in such an l length string, so $A_l \leq s^l$. Thus the right hand side is at most $k \cdot l_{\max}$ giving $(\sum_{i=1}^{r} s^{-|f(x^{(i)})|})^k \leq k \cdot l_{\max}$. Taking kth root and limit as $k \to \infty$, the result follows. □

Proof of Kraft's theorem. (We proved the theorem by labeling nodes in an s-ary tree and deleting subtrees from the tree to ensure that the code will be prefix. We showed that at least one leaf will remain at the end to label it for the last codeword. Here is another proof:)
Arrange the lengths in nondecreasing order, i.e., $l_1 \le \cdots \le l_r$. Define the numbers $w_1 := 0$ and for $j > 1$ let

$$w_j := \sum_{i=1}^{j-1} s^{l_j - l_i}.$$

This gives $w_j = s^{l_j} \sum_{i=1}^{j-1} s^{-l_i} < s^{l_j} \sum_{i=1}^{j} s^{-l_i} \le s^{l_j}$, thus the s-ary form of w_j has at most l_j digits. Let $f(x^{(j)})$ be the s-ary form of w_j "padded" with 0's at the beginning if necessary to make it have length exactly l_j for every j. This gives a code, we show it is prefix. Assume some $f(x^{(j)})$ is just the continuation of another $f(x^{(h)})$. (Then $l_j > l_h$, so $j > h$.) Thus cutting the last $l_j - l_h$ digits of $f(x^{(j)})$ we get $f(x^{(h)})$. This "cutting" belongs to division by $s^{l_j - l_h}$ (plus taking integer part), so this would mean $w_h = \left\lfloor \frac{w_j}{s^{l_j - l_h}} \right\rfloor = \left\lfloor s^{l_h} \sum_{i=1}^{j-1} s^{-l_i} \right\rfloor = s^{l_h} \sum_{i=1}^{h-1} s^{-l_i} + \left\lfloor s^{l_h} \sum_{i=h}^{j-1} s^{-l_i} \right\rfloor \ge w_h + 1$, a contradiction. □

Jensen's inequality and its consequences

We will need the following simple tool that has a lot of important consequences and that is often very useful when proving theorems in information theory. Recall from calculus the notion of convexity of a function first.

Def.: A function $g : [a, b] \to R$ is *convex* if for every $x, y \in [a, b]$ and $\lambda \in [0, 1]$ we have

$$g(\lambda x + (1 - \lambda)y) \le \lambda g(x) + (1 - \lambda)g(y).$$

We say that g is *strictly convex* if we have strict inequality whenever $0 < \lambda < 1$ and $x \ne y$.

Jensen's inequality: *Let g be a convex function and Z is a random variable. Then*

$$g\left(\mathbb{E}(Z)\right) \le \mathbb{E}\left(g(Z)\right)$$

Moreover, if g is strictly convex, then equality holds if and only if $Z = \mathbb{E}(Z)$ with probability 1, i.e. Z is not random but a constant.

Proof. A convex function is always above its tangent, thus

$$g(x) \ge g(x_0) + c \cdot (x - x_0)$$

for all x and x_0, where c is the rise of the tangent. It is true if $x_0 = \mathbb{E}\, Z$ and remains true if we put the random variable Z in the place of x

$$g(Z) \ge g(\mathbb{E}(Z)) + c \cdot (Z - \mathbb{E}(Z))$$

taking the expected value of both sides and using the linear property of expected value ($\mathbb{E}(aX + bY) = a\,\mathbb{E}\,X + b\,\mathbb{E}\,Y$)

$$\mathbb{E}\left(g(Z)\right) \ge g(\mathbb{E}(Z)) + c\,\mathbb{E}\left(Z - \mathbb{E}(Z)\right) = g(\mathbb{E}(Z))$$

□

Corollary 4 *If $P = (p_1, \dots, p_k)$ and $Q = (q_1, \dots, q_k)$ are two probability distributions, then*

$$\sum_{i=1}^{k} p_i \log \frac{p_i}{q_i} \ge 0,$$

and equality holds iff $q_i = p_i$ for every i.

Convention: To make the formulas above always meaningful, we use the "calculation rules" (for $a \ge 0, b > 0$) $0 \log \frac{0}{a} = 0 \log \frac{a}{0} = 0$ and $b \log \frac{b}{0} = +\infty, b \log \frac{0}{b} = -\infty$.

Proof. Let Z be a random variable such that it takes the value $\frac{q_i}{p_i}$ with probability p_i. The function $-\log x$ is convex, thus by Jensen's inequality

$$\sum_{i=1}^{k} p_i \log \frac{p_i}{q_i} = \sum_{i=1}^{k} p_i \left(-\log \frac{q_i}{p_i}\right) \ge -\log \left(\sum_{i=1}^{k} p_i \frac{q_i}{p_i}\right) = -\log \left(\sum_{i=1}^{k} q_i\right) = 0.$$

The condition of equality also follows from the corresponding condition in Jensen's inequality. □

Theorem 5

$$0 \leq H(X) \leq \log r,$$

where $r = |\mathcal{X}|$. $H(X) = 0$ *iff* X *takes a fix value with probability* 1, $H(X) = \log r$ *iff* X *is uniformly distributed.*

Proof. $0 \leq H(X)$ is clear by $\log p_i \leq 0$ for all $x^{(i)} \in \mathcal{X}$, since $0 \leq p_i \leq 1$. Equality can occur iff $p_i = 1$ for some i, then all other probabilities should be zero.
Applying the Corollary above to $q_i = 1/r \ \forall i$ gives $H(X) \leq \log r$ and also the condition for equality.