Data Compression - Seminar 4

October 29, 2013

Problem 1 (Uniquely decodable and instantaneous codes) *Let*

$$L = \sum_{i=1}^{m} p_i \ell_i^{100}$$

be the expected value of the 100th power of the word lengths associated with an encoding of the random variable X. Let $L_1 = \min L$ over all instantaneous codes; and let $L_2 = \min L$ over all uniquely decodable codes. What inequality relationship exists between L_1 and L_2?

Solution.

$$\begin{aligned} L &= \sum_{i=1}^{m} p_i \ell_i^{100} \\ L_1 &= \min\{L : \text{ instantaneous codes}\} \\ L_2 &= \min\{L : \text{ uniquely decodable codes}\} \end{aligned}$$

We have $L_1 \geq L2$, since all instantaneous codes are uniquely decodable. Any set of codeword lengths which achieve the minimum of L_2 will satisfy the Kraft inequality and hence we can construct an instantaneous code with the same codeword lengths, and hence the same L. Hence we have $L_1 \leq L_2$. From both these conditions, we must have $L_1 = L_2$. ■

Problem 2 (How many fingers has a Martian?) *Let*

$$S = \begin{pmatrix} S_1 & \dots & S_m \\ p_1 & \dots & p_m \end{pmatrix}.$$

The S_i's are encoded into strings from a D-symbol output alphabet in a uniquely decodable manner. If $m = 6$ and the codeword lengths are $(\ell_1, \ell_2, \dots, \ell_6) = (1, 1, 2, 3, 2, 3)$, find a good lower bound on D. You may wish to explain the title of the problem.

Solution. Uniquely decodable codes satisfy Kraft's inequality. Therefore

$$f(D) = D^{-1} + D^{-1} + D^{-2} + D^{-3} + D^{-2} + D^{-3} < 1.$$

We have $f(2) = 7/4 > 1$, hence $D > 2$. We have $f(3) = 26/27 < 1$. So a possible value of D is 3. Our counting system is base 10, probably because we have 10 fingers. Perhaps the Martians were using a base 3 representation because they have 3 fingers. (Maybe they are like Maine lobsters ?) ■

Problem 3 (Slackness in the Kraft inequality) *An instantaneous code has word lengths $\ell_1, \ell_2, \ldots, \ell_m$ which satisfy the strict inequality*

$$\sum_{i=1}^{m} D^{-\ell_i} < 1.$$

The code alphabet is $\mathcal{D} = \{0, 1, 2, \ldots, D-1\}$. Show that there exist arbitrarily long sequences of code symbols in $\mathcal{D}^*$ which cannot be decoded into sequences of codewords.

Solution. Instantaneous codes are pre x free codes, i.e., no codeword is a prefix of any other codeword. Let $n_{\max} = \max\{n_1, n_2, \ldots, n_q\}$. There are $D^{n_{\max}}$ sequences of length $n_{\max}$. Of these sequences, $D^{n_{max}-n_i}$ start with the i-th codeword. Because of the prefix condition no two sequences can start with the same codeword. Hence the total number of sequences which start with some codeword is $\sum_{i=1}^{q} D^{n_{\max}-n_i} = D^{n\max} \sum_{i=1}^{q} D^{-ni} < D^{n_{\max}}$. Hence there are sequences which do not start with any codeword. These and all longer sequences with these length nmax sequences as prefixes cannot be decoded. (This situation can be visualized with the aid of a tree.)

Alternatively, we can map codewords onto dyadic intervals on the real line corresponding to real numbers whose decimal expansions start with that codeword. Since the length of the interval for a codeword of length n_i is D^{-n_i}, and $\sum D^{-n_i} < 1$, there exists some interval(s) not used by any codeword. The binary sequences in these intervals do not begin with any codeword and hence cannot be decoded. ■

Problem 4 (Conditions for unique decodability) *Prove that a code C is uniquely decodable if (and only if) the extension*

$$C^k(x_1, x_2, \ldots, x_k) = C(x_1)C(x_2)\ldots C(x_k)$$

is a one-to-one mapping from $\mathcal{X}^k$ to D for every $k \geq 1$. (The only if part is obvious.)

Solution. If C^k is not one-to-one for some k, then C is not UD, since there exist two distinct sequences, $(x_1, \ldots, x_k)$ and $(x'_1, \ldots, x'_k)$ such that

$$C^k(x_1, \ldots, x_k) = C(x_1)\ldots C(x_k) = C(x'_1)\ldots C(x'_k) = C^k(x'_1, \ldots, x'_k).$$

Conversely, if C is not UD then by definition there exist distinct sequences of source symbols, $(x_1, \ldots, x_i)$ and $(y_1, \ldots, y_j)$, such that

$$C(x_1)C(x_2)\ldots C(xi) = C(y_1)C(y_2)\ldots C(y_j).$$

Concatenating the input sequences $(x_1, \ldots, x_i)$ and $(y_1, \ldots, y_j)$, we obtain

$$C(x_1) \ldots C(x_i) C(y_1) \ldots C(y_j) = C(y_1) \ldots C(y_j) C(x_1) \ldots C(x_i),$$

which shows that C^k is not one-to-one for $k = i + j$. ■

Problem 5 (The Sardinas-Patterson test for unique decodability) *A code is not uniquely decodable if and only if there exists a finite sequence of code symbols which can be resolved in two different ways into sequences of codewords. That is, a situation such as*

A_1	A_2	A_3	...	A_m
B_1	B_2	B_3	...	B_n

must occur where each Ai and each Bi is a codeword. Note that B1 must be a prefix of A1 with some resulting "dangling suffix". Each dangling suffix must in turn be either a prefix of a codeword or have another codeword as its prefix, resulting in another dangling suffix. Finally, the last dangling suffix in the sequence must also be a codeword. Thus one can set up a test for unique decodability (which is essentially the Sardinas-Patterson test [1]) in the following way: Construct a set S of all possible dangling suffixes. The code is uniquely decodable if and only if S contains no codeword.

(a) State the precise rules for building the set S .

(b) Suppose the codeword lengths are ℓ_i, $i = 1, 2, \ldots, m$. Find a good upper bound on the number of elements in the set S.

(c) Determine which of the following codes is uniquely decodable:

(i) $\{0, 10, 11\}$.

(ii) $\{0, 01, 11\}$.

(iii) $\{0, 01, 10\}$.

(iv) $\{0, 01\}$.

(v) $\{00, 01, 10, 11\}$.

(vi) $\{110, 11, 10\}$.

(vii) $\{110, 11, 100, 00, 10\}$.

(d) For each uniquely decodable code in part (c), construct, if possible, an infinite encoded sequence with a known starting point, such that it can be resolved into codewords in two different ways. (This illustrates that unique decodability does not imply nite decodability.) Prove that such a sequence cannot arise in a prefix code.

Solution. The proof of the Sardinas-Patterson test has two parts. In the first part, we will show that if there is a code string that has two different interpretations, then the code will fail the test. The simplest case is when the

concatenation of two codewords yields another codeword. In this case, S_2 will contain a codeword, and hence the test will fail. In general, the code is not uniquely decodable, iff there exists a string that admits two different parsings into codewords, e.g.

$$x_1x_2x_3x_4x_5x_6x_7x_8 = x_1x_2, x_3x_4x_5, x_6x_7x_8 = x_1x_2x_3x_4, x_5x_6x_7x_8.$$

In this case, S_2 will contain the string x_3x_4, S_3 will contain x_5, S_4 will contain $x_6x_7x_8$, which is a codeword. It is easy to see that this procedure will work for any string that has two different parsings into codewords; a formal proof is slightly more difficult and using induction. In the second part, we will show that if there is a codeword in one of the sets S_i, $i \geq 2$, then there exists a string with two diff erent possible interpretations, thus showing that the code is not uniquely decodeable. To do this, we essentially reverse the construction of the sets. We will not go into the details - the reader is referred to the original paper.

(a) Let S_1 be the original set of codewords. We construct S_{i+1} from Si as follows: A string y is in S_{i+1} iff there is a codeword x in S_1, such that xy is in S_i or if there exists a $z \in S_i$ such that zy is in S_1 (i.e., is a codeword). Then the code is uniquely decodable iff none of the S_i, $i \geq 2$ contains a codeword. Thus the set $S = \bigcup_{i \geq 2} S_i$.

(b) A simple upper bound can be obtained from the fact that all strings in the sets S_i have length less than $\ell_{\max}$, and therefore the maximum number of elements in S is less than $2^{\ell_{\max}}$.

(c) (i) $\{0, 10, 11\}$. This code is instantaneous and hence uniquely decodable.

(ii) $\{0, 01, 11\}$. This code is a suffix code (see problem 11). It is therefore uniquely decodable. The sets in the Sardinas-Patterson test are $S_1 = \{0, 01, 11\}$, $S_2 = \{1\} = S_3 = S_4 = \ldots$.

(iii) $\{0, 01, 10\}$. This code is not uniquely decodable. The sets in the test are $S1 = \{0, 01, 10\}$, $S_2 = \{1\}$, $S_3 = \{0\}$, ... Since 0 is codeword, this code fails the test. It is easy to see otherwise that the code is not UD - the string 010 has two valid parsings.

(iv) $\{0, 01\}$. This code is a suffix code and is therefore UD. The test produces sets $S_1 = \{0, 01\}$, $S_2 = \{1\}$, $S3 = \emptyset$.

(v) $\{00, 01, 10, 11\}$. This code is instantaneous and therefore UD.

(vi) $\{110, 11, 10\}$. This code is uniquely decodable, by the Sardinas-Patterson test, since $S_1 = \{110, 11, 10\}$, $S_2 = \{0\}$, $S3 = \emptyset$.

(vii) $\{110, 11, 100, 00, 10\}$. This code is UD, because by the Sardinas Patterson test, $S_1 = \{110, 11, 100, 00, 10\}$, $S_2 = \{0\}$, $S3 = \{0\}$, etc.

(d) We can produce infinite strings which can be decoded in two ways only for examples where the Sardinas Patterson test produces a repeating set. For example, in part (ii), the string $011111\ldots$ could be parsed either as

0,11,11,... or as 01,11,11, ... Similarly for (vii), the string $10000\ldots$ could be parsed as $100, 00, 00, \ldots$ or as $10, 00, 00, \ldots$. For the instantaneous codes, it is not possible to construct such a string, since we can decode as soon as we see a codeword string, and there is no way that we would need to wait to decode.

■

References

[1] A.A. Sardinas and G.W. Patterson. A necessary and sufficient condition for the unique decomposition of coded messages. In IRE Convention Record, Part 8, pages 104–108, 1953.