MATHEMATICAL TRIPOS PART II (2022–2023)
CODING AND CRYPTOGRAPHY
EXAMPLE SHEET 2 OF 4

1 In a Binary Symmetric Channel (BSC) we usually take the probability p of error to be less than $1/2$. Why do we not consider $1 \geq p > 1/2$? What if $p = 1/2$?

2 Suppose we connect two BSCs with error probabilities p and q in series or in parallel. How are the channel matrices related? (Note, in parallel, the answer should be a 4×4 matrix.)

3 Suppose we use eight hole tape with the standard paper tape code (i.e. the simple parity check code of length 8) and the probability that an error occurs at a particular place on the tape (i.e. a hole occurs where it should not or fails to occur where it should) is 10^{-4}. A program requires about 10 000 lines of tape (each line containing eight places) using the paper tape code. Using the Poisson approximation, direct calculation (possible with a hand calculator but really no advance on the Poisson method) or otherwise show that the probability that the tape will be accepted as error free by the decoder is less than .04%.

Suppose now that we use the Hamming scheme (making no use of the last place in each line). Explain why the program requires about 17 500 lines of tape but that any particular line will be correctly decoded with probability about $1 - (21 \times 10^{-8})$ and the probability that the entire program will be correctly decoded is better than 99.6%.

4 If there is a perfect e-error correcting binary code of length n, show that $V(n, e)$ divides 2^n. This condition is not sufficient for such a code to exist. We prove this by establishing the following results.

(i) Verify that $\frac{2^{90}}{V(90,2)} = 2^{78}$.

(ii) Suppose that C is a perfect 2-error correcting binary code of length 90 and size 2^{78}. Explain why we may suppose, without loss of generality, that the zero word $\mathbf{0} \in C$.

(iii) Let C be as in (ii) with $\mathbf{0} \in C$. Consider the set

$$X = \{\mathbf{x} \in \mathbb{F}_2^{90} : x_1 = 1,\ x_2 = 1, d(\mathbf{0}, \mathbf{x}) = 3\}.$$

Show that, corresponding to each $\mathbf{x} \in X$, we can find a unique $\mathbf{c}(\mathbf{x}) \in C$ such that $d(\mathbf{c}(\mathbf{x}), \mathbf{x}) = 2$. Show that $d(\mathbf{c}(\mathbf{x}), \mathbf{0}) = 5$.

(iv) Continuing with the argument of (iii), show that $c_i(\mathbf{x}) = 1$ whenever $x_i = 1$. If $\mathbf{y} \in X$, find the number of solutions to the equation $\mathbf{c}(\mathbf{x}) = \mathbf{c}(\mathbf{y})$ with $\mathbf{x} \in X$ and, by considering the number of elements of X, obtain a contradiction.

This result, obtained by Marcel Golay, shows that there is no perfect $(90, 2^{78})$-code. He found another case when $2^n/V(n, e)$ is an integer and there *does* exist an associated perfect code (now called the *Golay code.*)[1].

5 Determine the set of integers n for which the repetition code of length n is perfect.

[1] The deep connections between the Golay code and certain Mathieu groups (a class of sporadic finite simple groups) is beyond the scope of this course. See the great little book *From error correcting codes through sphere packings to simple groups* by Thomas Thompson (Carus Mathematical Monographs, 1983).

6 (i) Construct a $(7, 8, 4)$-code from Hamming's code.
(ii) Prove that if $\delta < n$ then $A(n, \delta) \leqslant 2A(n-1, \delta)$.
(iii) Prove that if δ is even then $A(n-1, \delta-1) = A(n, \delta)$.
(iv) Hence compute $A(6, 4)$.

7 Let C be an $[n, m, d]$-code. Show that

$$m(m-1)d \leqslant \sum \sum d(\mathbf{c}_i, \mathbf{c}_j) \leqslant \frac{1}{2} nm^2$$

where the sum is over all codewords $\mathbf{c}_i$ and $\mathbf{c}_j$ of C. Use this to give an upper bound on $A(n, d)$ in the case $n < 2d$.

8 Prove the *Singleton bound* for $A(n, d)$, namely,

(i) Suppose $n, d > 1$. If a binary $[n, m, d]$-code exists, then a binary $[n-1, m, d-1]$-code exists. Hence $A(n, d) \leqslant A(n-1, d-1)$.

(ii) Suppose $n, d \geqslant 1$. Then $A(n, d) \leqslant 2^{n-d+1}$.

9 Let $X_1, X_2, \ldots$ be a Bernouilli (memoryless) source with letters drawn from an alphabet $\mathcal{A}$. Let $c_n : \Sigma^n \to \{0, 1\}^*$ be an optimal binary code for $(X_1, \ldots, X_n)$ with word length random variable S_n.

(i) Use the Noiseless Coding theorem to show that $\frac{1}{n}\mathbb{E}(S_n) \to H(X_1)$ as $n \to \infty$.

(ii) Let $\varepsilon > 0$. By (i) there exists $N \geqslant 1$ with $\mathbb{E}(S_N) < N(H(X_1) + \varepsilon)$. By considering the sets

$$A_n = \{x \in \Sigma^n : c_N^*(x) \text{ has length } \leqslant n(H(X_1) + 2\varepsilon)\}$$

for n a multiple of N, show that the source is reliably encodable at rate $H(X_1) + 2\varepsilon$.

What does this tell you about the information rate of the source? (In lectures we showed that the information rate is $H(X_1)$ using the AEP - see [GP, 2.3 and 2.4].)

10 Consider two DMCs of capacity C_1 and C_2 with each having input alphabet Σ_1 and output alphabet Σ_2. Connecting in parallel gives the *product channel* with input alphabet $\Sigma_1 \times \Sigma_1$, output alphabet $\Sigma_2 \times \Sigma_2$, and channel probabilities given by

$$\mathbb{P}(y_1 y_2 \text{ received } | x_1 x_2 \text{ sent}) = \mathbb{P}(y_1 \text{ received } | x_1 \text{ sent}) \mathbb{P}(y_2 \text{ received } | x_2 \text{ sent}).$$

Show that the product channel has capacity $C = C_1 + C_2$.

11 Show that the binary channel with channel matrix

$$\begin{pmatrix} 1 & 0 \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

has capacity $\log 5 - 2$.

12 Players A and B play a (best of) 5 set tennis match. Let X be the number of sets won by A, and let Y be the total number of sets played. Assuming that the players are equally matched and the outcome of each set is independent, compute the conditional entropies $H(X|Y)$, $H(Y|X)$ and the mutual information $I(X;Y)$.

Further Problems

13 If a channel matrix, with output alphabet of size n, is such that the set of entries in any row is the set $\{p_1, \dots, p_n\}$, and the set of entries in each column is the same, show that its information capacity C is given by

$$C = \log n + \sum_{i=1}^{n} p_i \log p_i.$$

Hence show that the capacity of the channel that has matrix

$$\begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{6} & \frac{1}{6} \\ \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}$$

is given by $C = \log 2^{5/3} - \log 3$. This is taken from Welsh's book; Welsh credits it to Shannon (1948).

14 If you look at the inner title page of almost any book published between 1974 and 2007, you will find its International Standard Book Number (ISBN-10). The ISBN-10 uses single digits selected from $0, 1, \dots, 8, 9$ and X representing 10. Each ISBN-10 consists of nine such digits a_1, a_2, $\dots$, a_9 followed by a single check digit a_{10} chosen so that

$$(*) \qquad 10a_1 + 9a_2 + \cdots + 2a_9 + a_{10} \equiv 0 \pmod{11}.$$

(In more sophisticated language, our code C consists of those elements $\mathbf{a} \in \mathbb{F}_{11}^{10}$ such that $\sum_{j=1}^{10}(11 - j)a_j = 0$.)

(i) Find a couple of books[2] and check that $(*)$ holds for their ISBNs.

(ii) Show that $(*)$ will not work if you make a mistake in writing down one digit of an ISBN.

(iii) Show that $(*)$ may fail to detect two errors.

(iv) Show that $(*)$ will not work if you interchange two distinct adjacent digits (a transposition error).

(v) Does (iv) remain true if we remove the word 'adjacent' ? Errors of type (ii) and (iv) are the most common in typing.

In communication between publishers and booksellers, both sides are anxious that errors should be detected but would prefer the other side to query errors rather than to guess what the error might have been.

(vi) Since the ISBN contained information such as the name of the publisher, only a small proportion of possible ISBNs could be used[3] and the system described above started to 'run out of numbers'. A new system was introduced which is compatible with the system used to label most consumer goods.

After January 2007, the appropriate code became a 13 digit ISBN-13 number $x_1x_2 \dots x_{13}$ with each digit selected from 0, 1, $\dots$, 8, 9 and the check digit x_{13} computed by using the formula

$$x_{13} \equiv -(x_1 + 3x_2 + x_3 + 3x_4 + \cdots + x_{11} + 3x_{12}) \pmod{10}.$$

Show that we can detect single errors. Give an example to show that we cannot detect all transpositions.

[2] try a place called the 'College Library' (ask the Porters where it is).

[3] The same problem occurs with telephone numbers. If we use the Continent, Country, Town, Subscriber system we will need longer numbers than if we just numbered each member of the human race.

15 In a horse race at Royal Basket are entered m horses. The probability that the i^{th} horse wins is p_i. The odds offered on each horse are a_i–for–1, *i.e.* a bet of x pounds on the i^{th} horse will yield $a_i x$ pounds if the horse wins, and nothing otherwise. A gambler bets a proportion b_i of his wealth on horse i, with $\sum_{i=1}^m b_i = 1$. She seeks to maximise $W = \sum_{i=1}^m p_i \log(a_i b_i)$. Suggest a motivation for this choice. Solve to find the b_i that maximise W. Show that, in the case when all odds are the same, this maximum and the entropy $H(p_1, \ldots, p_m)$ sum to a constant.

16 Consider the situation described in Chapter 10 and in particular Theorem (10.2).

(i) Show that, for the situation described you should not bet if $up \leqslant 1$ and should take

$$w = \frac{up - 1}{u - 1}$$

if $up > 1$.

(ii) Write $q = (1 - p)$. Show that, if $up > 1$ and we choose the optimum w,

$$\mathbb{E} \log Y = p \log p + q \log q + \log u - q \log(u - 1).$$

(iii) Show that, if you bet less than the optimal proportion, your fortune will still tend to increase but more slowly, but, if you bet more than some proportion w_1, your fortune will decrease. Write down the equation for w_1.

The moral of this story is that if you use the Kelly criterion veer on the side of underbetting.[4]

SM, Lent Term 2023
Comments on and corrections to this sheet may be emailed to `sm137@cam.ac.uk`

[4] For more on proportional gambling or the Kelly criterion, see 2.6 of Körner's book *Naive decision making* or Chapter 6 of [CT]. Kelly's original 9-page paper from 1956, *A new interpretation of information rate* is freely available and is a model of clarity.