MATHEMATICAL TRIPOS PART II (2019–2020)
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 BSC's 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 A Binary Symmetric Channel with error probability $p = \frac{1}{3}$ is used to send codewords 1100, 0110, 0001, 1111 with probabilities $\frac{1}{4}, \frac{1}{2}, \frac{1}{12}, \frac{1}{6}$. How would you decode 1001 using (i) ideal observer decoding, or (ii) maximum likelihood decoding?

4 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%.

5 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].

[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 (I kid you not) Thomas Thompson of Walla Walla College (Carus Mathematical Monographs, 1983).

6 Determine the set of integers n for which the repetition code of length n is perfect. Show that the repetition code of length n is perfect if and only if n is odd.

7 (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 $A(6, 4)$.

8 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 $n < 2d$.

9 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 $A(n, d) \leqslant 2^{n-d+1}$.

10 (i) Show that $H(X|Y) \geqslant 0$ with equality if and only if X is a function of Y. (ii) Give an example where $H(X|Y = y) > H(X)$, even though $H(X|Y) \leq H(X)$.

11 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

$$\mathbb{P}(y_1y_2 \text{ received } | x_1x_2 \text{ sent}) = \mathbb{P}(y_1 \text{ received } | x_1 \text{ sent})\mathbb{P}(y_2 \text{ received } | x_2 \text{ sent}).$$

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

12 Show that the capacity of the DMC with channel matrix

$$\begin{pmatrix} 1-\alpha-\beta & \alpha & \beta \\ \alpha & 1-\alpha-\beta & \beta \end{pmatrix}$$

is $C = (1-\beta)(1-\log(1-\beta)) + (1-\alpha-\beta)\log(1-\alpha-\beta) + \alpha\log\alpha$.

Further Problems

13 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 $I(X;Y)$.

14 Codewords 00 and 11 are sent with equal probability through a BSC with error probability p. Compute the mutual information between the codeword sent and the first digit received as output. Show that the extra mutual information to accrue on receipt of the second digit is $H(2p(1-p)) - H(p)$ bits.

15 In an examination each candidate is asked to write down a Candidate Number of the form $2234A$, $2235B$, $2236C$,... (the eleven possible letters are repeated cyclically) and a Desk Number. (Thus candidate 0004 sitting at desk 425 writes down $0004D - -425$.) The first four numbers in the Candidate Number identify the candidate uniquely. Show that if the candidate makes one error in the Candidate Number then that error can be detected without using the Desk Number. Would this be true if there were nine possible letters repeated cyclically? Would this be true if there were twelve possible letters repeated cyclically? Give reasons.

Show that if we combine the Candidate Number and the Desk Number the combined code is 1-error correcting.

16 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, ..., 8, 9 and X representing 10. Each ISBN-10 consists of nine such digits a_1, a_2, ..., 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 $x_1x_2\ldots x_{13}$ with each digit selected from 0, 1, ..., 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.

SM, Lent Term 2020
Comments on and corrections to this sheet may be emailed to `sm@dpmms.cam.ac.uk`

[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.