COMMUNICATION THEORY: Example Sheet 1

Information sources and coding

Decipherability, prefix-free coding, the Kraft inequality, Shannon's NC theorem

In this part of the course, a *code* (also *coding* or *encoding*) is a map

$$f:\ u \in I \mapsto f(u) \in J^* = \cup_{n\geq 1} J^n,$$

where $I(=I_m)=\{1,\ldots,m\}$ is a source alphabet, $J(=J_a)=\{0,\ldots,a-1\}$ is an encoder alphabet and J^* the set of finite words, or strings, from J (J^n is the set of strings $x^{(n)}=x_1\ldots x_n$ of length n). The strings $x\in J^*$ that are images, under f, of symbols $u\in I$ are called *codewords* (in code f). A source *message* (of length n) is a sequence $u_1\ldots u_n$ of symbols $u_j\in I$; it is encoded as a concatenation of the codewords $f(u_1)\ldots f(u_n)$.

A code f is called *decipherable* if any string $x\in J^*$ may be written as a concatenation of at most one collection of codewords. A code f is called *prefix-free* if no codeword $x\in J^*$ may occur as a prefix in another codeword y (i.e., the representation $y=xx'$ is impossible for any pair of codewords x,y). Any prefix-free code is decipherable, but not vice versa. The Kraft inequality

$$\sum_{i=1}^{m} a^{-s_i} \leq 1$$

is sufficient for the existence of a prefix-free and necessary for the existence of a decipherable code, with codeword-lengths $s_1,\ldots,s_m$.

Dealing with *random* symbols $u\in I$ appearing with probabilities (or frequencies) $p(u)$, one is interested in *optimal* (decipherable) codes f for which $\mathbf{E}S$, the expected value of the random codeword-length, is minimal. An optimal code that is prefix-free was constructed by Huffman; in the case of binary coding ($a=2$ and $J=\{0,1\}$) Huffman's codes are related to binary trees. Another class of codes is formed by the so-called Shannon-Fano codes; the Shannon-Fano codes are simple in implementation and 'close' to optimal. Their construction is based on Shannon's noiseless coding theorem.

1. Determine Huffman's binary coding when the distribution over the original alphabet is:

(a) (0.3, 0.2, 0.2, 0.2, 0.1), (b) (1/4, 1/4, 1/4, 1/8, 1/8),
(c) (1/3, 1/5, 1/5, 2/15, 2/15), (d) (0.1, 0.1, 0.1, 0.15, 0.26, 0.29).

2. A drawback of Huffman's encoding is that the codeword-lengths are complicated functions of the symbol probabilities $p_1,\ldots,p_m$. However, some bounds are available. Suppose that $p_1\geq p_2\geq\ldots\geq p_m$. Prove that in *any* binary Huffman ecoding
(a) if $p_1>2/5$ then letter 1 must be encoded by a codeword of length 1,

(b) if $p_1<1/3$ then letter 1 must be encoded by a codeword of length ≥ 2.

3. A Shannon–Fano code is in general not optimal. However, it is 'not much' longer than Huffman's. Prove that if $S_{\rm SF}$ is the Shannon–Fano codeword–length then for any $r=1,2,\ldots$ and any decipherable code f^* with the codeword–length S^*,

$$\mathbf{P}\big(S^*\leq S_{\rm SF}-r\big)\leq a^{-(r-1)}$$

Entropy, conditional entropy, mutual entropy

Given an event A with probability $\mathbf{P}(A)$, the *information* gained from the fact that A has occurred is defined as $i(A)=-\log_2\mathbf{P}(A)$. Let X be a random variable (r.v.) taking values $\{1,\ldots,m\}$, with probabilities $p_1,\ldots,p_m$. The *entropy* $h(X)$ is defined as the expected amount of information gained from observing X:

$$h(X)=-\sum_i p_i\log_2 p_i.$$

Here and below we set $0\cdot\log_2 0=0$. It is clear that the entropy $h(X)$ depends in fact on the probability distribution: $h(X)=h(p_1,\ldots,p_m)$.

The *conditional* entropy, $h(X|Y)$, of a r.v. X, given r.v. Y, is defined as the expected amount of information gained from observing X given that a value of Y is known:

$$h(X|Y)=-\sum_{i,j} p_{X,Y}(i,j)\log\ p_{X|Y}(i|j).$$

Here, $p_{X,Y}(i,j)$ is the joint probability $\mathbf{P}(X=i,Y=j)$ and $P_{X|Y}(i|j)$ the conditional probability $\mathbf{P}(X=i|Y=j)$. It is easy to check that $h(X|Y)=h(X,Y)-h(Y)$, which yields

$$0\leq h(X|Y)\leq h(X),$$

with the LH equality iff $X=\phi(Y)$ and the RH equality iff X and Y are independent, and

$$h(X,Y|Z)\leq h(X|Z)+h(Y|Z),$$

with the equality iff X and Y are conditionally independent given Z: $\mathbf{P}(X=x,Y=y|Z=z)=\mathbf{P}(X=x|Z=z)\mathbf{P}(Y=y|Z=z)$.

The mutual entropy (or mutual information), $i(X,Y)$, between r.v.'s X and Y is defined by

$$i(X,Y)=\sum_{i,j} p_{X,Y}(i,j)\log\frac{p_{X,Y}(i,j)}{p_X(i)p_Y(j)}.$$

In other words, $i(X,Y)=h(X)+h(Y)-h(X,Y)$. You see that

$$i(X,Y)\geq 0,$$

with the equality iff X and Y are independent.

4. Show that the quantity

$$\rho(X,Y)=h(X|Y)+h(Y|X)$$

obeys

$$\rho(X,Y) = h(X) + h(Y) - 2i(X,Y) = h(X,Y) - i(X,Y) = 2h(X,Y) - h(X) - h(Y).$$

Prove that ρ has the following properties: a) $\rho(X,Y) = \rho(Y,X) \geq 0$, b) $\rho(X,Y) + \rho(Y,Z) \geq \rho(X,Z)$. Show that c) $\rho(X,Y) = 0$ iff X and Y are functions of each other. Also show that if X' and X are functions of each other then $\rho(X,Y) = \rho(X',Y)$. Hence, ρ may be considered as a *metric* on the set of the random variables X, considered modulo the equivalence: $X \sim X'$ iff X and X' are functions of each other. [Property a) guarantees the symmetry and b) is the triangle inequality.]

[**Hint** (to bound b)): *Prove a stronger inequality* $h(X,Z) \leq h(X,Y) + h(Y,Z) - h(Y)$.]

5. Write $h(\mathbf{p}) := -\sum_1^m p_j \log p_j$ for a probability 'vector' $\mathbf{p} = (p_1 \dots, p_m)^T$.

a) Show that $h(P\mathbf{p}) \geq h(\mathbf{p})$ if P is a doubly stochastic matrix (i.e. a square matrix of non-negative elements for which all row and column sums are unity).

[**Hint:** *Use the fact that, for any non-negative* λ_i, c_i *such that* $\sum_1^m \lambda_i = 1$, $\log(\lambda_1 c_1 + \dots + \lambda_m c_m) \geq \sum_1^m \lambda_i \log c_i$.]

Show that the two are equal if and only if P is a permutation matrix.

b) Show that $h(\mathbf{p}) \geq -\sum_{j=1}^m \sum_{k=1}^m p_j P_{j,k} \log P_{j,k}$ if P is a stochastic matrics (a square matrix of non-negative elements where all row sums are unity) and $\mathbf{p}$ is an invariant vector of P: $\mathbf{p}P = \mathbf{p}$.

[**Hint:** *Use Gibbs' inequality.*]

6. The sequence of random variables $\{X_j : j = 1,2,\dots\}$ forms a Markov chain with a finite state space.

a) Quoting standard properties of conditional entropy, show that

$$h(X_j|X_{j-1}) \leq h(X_j|X_{j-2}) \leq 2h(X_j|X_{j-1}).$$

b) Show that the mutual information $i(X_m, X_n)$ is non-decreasing in m and non-increasing in n, $1 \leq m \leq n$.

Shannon's FC theorem. Information rates of Bernoulli and Markov sources

Here, one deals with random sources emitting a sequence of random symbols U_1, ..., U_n, ... from alphabet I. The random string $(U_1, \dots, U_n)$ of length n is denoted by $U^{(n)}$; it takes values of $u^{(n)}$ with probabilities $p(u^{(n)}) := \mathsf{P}(U^{(n)} = u^{(n)})$. The two examples are (i) a Bernoulli source, where $U_1, U_2, \dots$ is a sequence of independent, identically distributed random symbols, and (ii) a Markov source, where $U_1, U_2, \dots$, is a Markov chain.

A source is said to be *reliably encodable* at rate $R > 0$ if there exists a sequence of sets $A_n \subseteq I^n$ such that $\sharp\, A_n \leq 2^{nR}$ and $\lim_{n\to\infty} \mathsf{P}(U^{(n)} \in A_n) = 1$. The *information rate* of a source is defined as

$$H = \inf [R: \ R \text{ is reliable }].$$

A way of calculating H is provided by Shannon's First coding theorem. Namely, define a r.v. $\xi^{(n)}$ by

$$\xi^{(n)} = p(u^{(n)}), \text{ if } U^{(n)} = u^{(n)}.$$

If there exists a nonrandom limit $\lim_{n\to\infty} \frac{1}{n} \xi^{(n)}$, then this limit equals H. The existence of the limit is usually derived from the Law of large numbers and is guaranteed for a Bernoulli source and for a Markov source if, e.g., it is irreducible and aperiodic. Correspondingly, for a Bernoulli source

$$H = -\sum_{i=1}^m p_i \log p_i, \text{ where } p_i = \mathsf{P}(U_n = i),$$

and for an irreducible and aperiodic Markov source

$$H = -\sum_{i=1}^m w_i \sum_{j=1}^m P(i,j) \log P(i,j),$$

where $\{w_i\}$ is the (unique) invariant distribution of the Markov chain and $P(i,j)$ is the transition probability. In other words, for a Bernoulli source, $H = h(U_n)$ and for an irreducible and aperiodic Markov source $H = \lim_{n\to\infty} h(U_{n+1}|U_n)$ (if the source is stationary, then $H = h(U_{n+1}|U_n)$). The condition that the Markov source is irreducible and aperiodic is not necessary and in some cases may be weakened.

7. Consider a source with letters chosen from an alphabet of size $a + b$, for which the message strings are constrained by the condition that no two letters of A should ever occur consecutively, where A is a subset of the alphabet of size a.

a) Suppose the message follows a Markov chain, all characters which are permitted at a given place being equally likely. Show that this source has information rate

$$H = \frac{a \log b + (a+b)\log(a+b)}{2a+b}.$$

b) By solving a recurrence relation, or otherwise, find how many strings of length n satisfy the constraint that no two letters of A occur consecutively. Suppose these strings are equally likely and let $n \to \infty$. Show that the limiting information rate becomes

$$H = \log\left(\frac{b + \sqrt{b^2 + 4ab}}{2}\right).$$

8. Let $\{U_j : j = 1,2,\dots\}$ be an irreducible and aperiodic Markov chain with a finite state space. Given $n \geq 1$ and $\alpha \in (0,1)$, order the strings $u^{(n)}$ according to their probabilities $\left(\mathsf{P}(U^{(n)} = u_1^{(n)}) \geq \mathsf{P}(U^{(n)} = u_2^{(n)}) \geq \dots\right)$ and select them in the order until the probability of the remainig set becomes $\leq 1 - \alpha$. Let $M_n(\alpha)$ denote the number of the selected strings. Prove that $\lim_{n\to\infty} \frac{1}{n} \log M_n(\alpha) = H$, the information rate of the source,

a) in the case where the rows of the transition probability matrix P are all equal (i.e., $\{U_j\}$ is a Bernoulli sequence),

b) in the case where the rows of P are permutations of each other,

c*) in a general case.

COMMUNICATION THEORY: Example Sheet 2

Channel capacity

A *channel* is defined by specifying a conditional probability distribution $\mathsf{P}(y^{(N)}|x^{(N)})$, where $x^{(N)}$ is a (binary) word (or string) of length N sent to the input and $y^{(N)}$ a word received on the output. [The reason for changing from n to N is duscussed below.] More precisely, one should specify these distributions for each N; in the case of a *memoryless* channel,

$$\mathsf{P}(y^{(N)}|x^{(N)}) = \prod_{j=1}^{N} P(y_j|x_j).$$

A memoryless channel is described by a channel *matrix* P whose x,y-entry is $P(y|x)$, $x = 0,1$, $y \in \widetilde{J}$ (the output alphabet). If $\widetilde{J} = \{0,1\}$ the channel is called binary; in this case the channel matrix is 2×2. A memoryless binary channel is called *symmetric* if $P(1|0) = P(0|1)$ (and hence $P(1|1) = P(0|0)$).

If $y^{(N)} \neq x^{(N)}$, there is an error occurred while string $x^{(N)}$ was sent through the channel, and the corresponding probability equals

$$\mathsf{P}\Big(\{y^{(N)} \neq x^{(N)}\}|x^{(N)}\Big) = \sum_{y^{(N)} \neq x^{(N)}} \mathsf{P}(y^{(N)}|x^{(N)}).$$

The channel *capacity* C is defined as the supremum $\sup \overline{R}$ of the reliable *transmission* rates $\overline{R} \in (0,1)$. The definition of a reliable transmission rate (for a given channel) is that for each n you can find a code $f\colon \{0,1\}^n \to \{0,1\}^N$ and a decoding rule $\widehat{f}\colon \{0,1\}^N \to \{0,1\}^n$, with $N = [\overline{R}^{-1} n]$ such that

$$\lim_{n\to\infty} 2^{-n} \sum_{u^{(n)} \in \{0,1\}^n} \mathsf{P}\Big(\{y^{(N)} : \widehat{f}(y^{(N)}) \neq u^{(n)}\} \mid f(u^{(n)})\Big) = 0.$$

The key moment in this definition is that you are allowed to increase the length n of the source string by factor $\overline{R}^{-1} > 1$; by introducing a redundancy in your encoding you might be able to cope with the errors in the channel. A reliable transmission rate is the one for which such a construction is possible. Another feature is the factor 2^{-n} in front of the sum $\sum_{u^{(n)}}$: it means that you consider *equidistributed* source strings $U^{(n)}$.

A way of calculating the channel capacity (together with optimal coding and decoding) is provided by Shannon's Second coding theorem. In particular, for a memoryless channel,

$$C = \sup_{p_X} i(X,Y).$$

Here, $i(X,Y)$ is the mutual entropy between the random symbol X on the input and the corresponding symbol Y on the output of the cahnnel; the supremum is over all

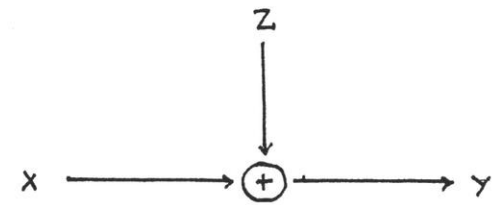
probability distributions p_X of the input symbol. For the memoryless binary symmetric channel this yields

$$C = 1 - h(p, 1-p)$$

where $p = P(1|0) = P(0|1)$ is the symbol error probability and $h(p,1-p) = -p\log p - (1-p)\log(1-p)$ is the binary entropy. Here, the channel capacity is achieved while using the equidistributed input symbols 0 and 1 ($p_X(0) = p_X(1) = 1/2$).

1. One is given a memoryless channel with transition probabilities $p(y|x)$ and channel capacity $C = \max_{p(x)} i(X,Y)$. A helpful statistician preprocesses the output by forming $Y' = g(Y)$: he claims that this will strictly improve the capacity.
   (a) Show that he is wrong.
   (b) Under what condition does he not strictly decrease the capacity?

2. Find the capacity of the following memoryless channel:



where an additive noise Z takes values 0 and a with probability $1/2$, a is a given real number. The input alphabet is $\{0,1\}$ and Z is independent of X. Why does the capacity depend on a?

3. A channel has binary input and output alphabets and transition probabilities $p(y|x)$ given by the following matrix

$$\begin{pmatrix} 1 & 0 \\ 1/2 & 1/2 \end{pmatrix}.$$

Find the capacity of the channel and the maximizing input probability distribution.

4. Bits are transmitted along a communication channel. With probability λ a bit may be inverted and with probability μ it may be rendered illegible. The fates of successive bits are independent. Determine the optimal coding for, and the capacity of, the cahnnel.

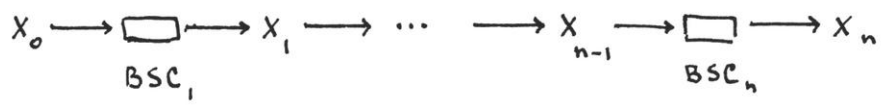
5. Let X,Y be jointly distributed discrete random variables. Define the conditional entropy $h(X|Y)$. A message X is transmitted through a first channel, and the output Y of this channel is then transmitted through a second channel to give a final output Z. Thus the discrete random variables X,Y,Z are such that X and Z are independent, conditional on the value of Y. Show that

$$h(X|Y) \leq h(X|Z),$$

and determine conditions under which there is equality in this relation.

6. The input and output of a discrete-time channel are both expressed in an alphabet whose letters are the residue classes of integers modulo r, where r is fixed. The transmitted letter $[x]$ is received as $[j+x]$ with probability p_j, where x and j are integers and $[c]$ denotes the residue class of c mod r. Calculate the capacity of the channel.

7. Find the error probability of a cascade of n identical independent binary symmetric channels (BSC)



each with the error probability p. Show that the capacity of the casacade tends to zero as $n \to \infty$.

8. A spy sends messages to his contact as follows. Each hour either he does not telephone, or he telephones and allows the telephone to ring a certain number of times – not more than N, for fear of detection. His contact does not answer, but merely notes whether or not the telephone rings, and, if so, how many times. Because of deficiencies in the telephone system, calls may fail to be properly connected; correct connection has probability p, where $0 < p < 1$, and is independent for distinct calls, but the spy has no means of knowing which calls reach his contact. If connection is made, then the number of rings is transmitted correctly. The probability of a false conection when no call is made may be neglected. Write down the channel matrix for this channel and calculate the capacity explicitly. Determine a condition on N in terms of p which will imply, with optimal coding, that the spy will always telephone.

9. Suppose one has two independent memoryless discrete channels, with capacities C_1, C_2 bits/sec. Prove, or provide a counterexample to, each of the following claims about the capacity C of a compound channel formed as stated.

(a) If the channels are in series, with the output from one being fed into the other with no further coding, then $C = \min(C_1, C_2)$.

(b) If the channels are used in parallel in the sense that at every second a symbol is transmitted through channel 1 (from its input alphabet) and a symbol through chanel 2 (from its input alphabet); each channel thus emits one symbol each second. Then $C = C_1 + C_2$.

(c) If the channels have the same input alphabet and at each second a symbol is chosen and sent simultaneously down both channels, then $C = \max(C_1, C_2)$.

(d) If channel i has matrix Π_i and the compound channel has

$$\Pi = \begin{pmatrix} \Pi_1 & 0 \\ 0 & \Pi_2 \end{pmatrix}$$

then C is given by $2^C = 2^{C_1} + 2^{C_2}$. To what mode of operation does this correspond?

Lent 1996 YMS
E-mail: yms@statslab.cam.ac.uk

COMMUNICATION THEORY: Example Sheet 3

The Hamming space is the set $\{0,1\}^N$ of all binary words of lenght N, endowed with the Hamming distance

$$d(x^{(N)}, y^{(N)}) = \text{ the number of digits } i \text{ with } x_i \neq y_i.$$

A ball of radius R about a word $x^{(N)}$ in the Hamming distance is the set $\{y^{(N)} : d(x^{(N)}, y^{(N)}) \leq R\}$.

The Hamming space is closed under an operation of addition modulo 2:

$$x^{(N)} + y^{(N)} = (x_1 + y_1) \text{mod } 2 \ \ldots \ x_N + y_N \text{mod } 2$$

[Here, $1 + 1 = 0$ mod 2.] More precisely, the Hamming space is a commutative group with the componentwise addition and with the zero codeword $\mathbf{0} = 0 \ldots 0$ playing the role of the zero of the group. Each element of this group is opposite to itself: $x^{(N)} + x'^{(N)} = \mathbf{0}$ iff $x^{(N)} = x'^{(N)}$.

Henceforth, all operations over the binary words are understood mod 2.

A *code* of length N is a subset $\mathcal{X}_N \subseteq \{0,1\}^N$; the words $x^{(N)} \in \mathcal{X}_N$ are called the codewords. The number of words in a code, $\mathcal{X}_N$, is called the size of the code, and a code of lenght N and size r is called an $[N, r]$ code. The information rate of an $[N, r]$ code is defined as $\rho = \dfrac{\log r}{n}$. Two codes are called equivalent iff they are obtained from each other by a permutation of digits in the binary words.

Another important parameter of a code is the minimum distance (or briefly, the distance). A code has distance δ if

$$\delta = \min [d(x^{(N)}, y^{(N)}) : \ x^{(N)}, y^{(N)} \in \mathcal{X}_N, \ x^{(N)} \neq y^{(N)}].$$

A code $\mathcal{X}_N$ is called *linear* if $\mathcal{X}_N$ contains, with each pair of codewords $x^{(N)}$ and $y^{(N)}$, their sum $x^{(N)} + y^{(N)}$. The maximal number of linearly independent words from a linear code is called the rank of the code. [A collection of words is called linearly independent if the sum of any subset of words from the collection gives a non-zero word.] A linear code of length N and rank k is called a (linear) (N, k) code. An (N, k) code of minimum distance δ is called an (N, k, δ) code.

Let $\mathcal{X}_N$ be a linear (N, k) code. Any collection of k linearly independent codewords from $\mathcal{X}_N$ is called a basis in $\mathcal{X}_N$. All codewords are obtained as linear combinations of words from a basis. A $k \times n$ matrix G whose rows are basis vectors is called a generating matrix of code $\mathcal{X}_N$. A parity check matrix of code $\mathcal{X}_N$ is an $n \times (n - k)$ matrix with linearly independent columns such that $\mathcal{X} = \{x^{(N)} : \ x^{(N)} H = \mathbf{0}\}$.

A linear code is called cyclic if, together with each codeword $x^{(N)} = x_1 \ \ldots \ x_N$, it contains its cyclic shift $\Pi x^{(N)} = x_N x_1 \ \ldots \ x_{N-1}$.

1. (a) A code is said to be D-error detecting if making up to D changes in a codeword does not lead to another codeword. What is the minimum distance of a D-detecting code? Prove that if the distance δ of an $[N, r]$ code $\mathcal{X}$ is an odd number then the code may be extended to an $[N + 1, r]$ code $\mathcal{X}^+$ with distance $\delta + 1$.

(b) A code $\mathcal{X}$ is said to be E-error correcting if making up to E changes in any codeword $x \in \mathcal{X}$ leads to a word y such that x is still closer to y than any other codeword. Prove that $\mathcal{X}$ is an E-error correcting code iff the balls of radius E about the codewords are disjoint. Hence show that $\mathcal{X}$ can be extended to a code $\mathcal{X}^+$ that detects $2E + 1$ errors.

(c) An E-errror correcting code $\mathcal{X}_N$ is called perfect if the balls of radius E about codewords $x \in \mathcal{X}_N$ cover the whole Hamming space $\{0,1\}^N$. Show that the distance of a perfect code is an odd number.

2. (a) The parity-check code $\mathcal{X}_N$ is defined as the set of binary words $x^{(N)} = x_1 \ \ldots \ x_N$ with $\sum_{i=1}^N x_i = 0$. Prove that this code is linear and find its rank and minimal distance. Write its generating and parity-check matrix. What is the information rate of this code?

(b) Let $\mathcal{X}$ be a linear (N, k, δ) code. Let G and H be, respectively, the generating and parity-check matrices of $\mathcal{X}$. The parity-check extention of $\mathcal{X}$ is a code $\mathcal{X}^+$ of length $N + 1$ obtained by adding, to each codeword $x \in \mathcal{X}_N$, the symbol $x_{N+1} = \sum_{i=1}^N x_i$ so that the sum $\sum_{i=1}^{N+1} x_i$ is zero. Prove that $\mathcal{X}^+$ is a linear code and find its rank and minimal distance. How are the information rates and generating and parity-check matrices of $\mathcal{X}$ and $\mathcal{X}^+$ related?

(c) The truncation $\mathcal{X}^-$ of an (N, k, δ) code $\mathcal{X}$, with the generating and parity-check matrices G and H, is defined as a linear code of length $N - 1$ obtained by omitting the last symbol of each codeword $x \in \mathcal{X}$. Supose that code $\mathcal{X}$ has $\delta \geq 2$. Prove that $\mathcal{X}^-$ is linear and find its rank and generating and parity-check matrices. Show that the minimal distance is $\geq \delta - 1$.

(d) Let $\mathcal{X}$ be a linear (N, k, d) code with the generating and parity-check matrices G and H. The m-repetition extention of $\mathcal{X}$ is a code $\mathcal{X}^{\text{re}}_{(m)}$ of length Nm obtained by repeating each codeword $x \in \mathcal{X}$ m times. Prove that $\mathcal{X}^{\text{re}}_{(m)}$ is a linear code and find its rank and minimal distance. How are the information rates and generating and parity-check matrices of $\mathcal{X}$ and $\mathcal{X}^{\text{re}}_{(m)}$ related?

3. What is the number of codewords in a linear (N, k)-code? What is the number of different bases in it? Calculate the last number for $k = 4$. List all bases for $k = 2$ and $k = 3$.

Show that the subset of a linear code consisting of all words of even weight is a linear code.

Prove that if there exists a linear (N, k, d)–code then there exists a linear (N, k, d)–code with codewords of even weight.

4. A dual code of a linear (N, k) code $\mathcal{X}$ is defined as the set of the words $y = y_1 \ \ldots \ y_N$

such that the dot-product

$$\langle y \cdot x \rangle = \sum_{i=1}^{N} y_i x_i = 0 \quad \text{for each} \quad x = x_1 \ldots x_N \in \mathcal{X}_N.$$

Prove that an $N \times (N-k)$ matrix H is a parity-check matrix of code $\mathcal{X}$ iff H^{T} is a generator for the dual code. Hence, derive that G and H are generating and parity-check matrices, respectively, for a linear code iff

(i) the rows of G are linearly independent,

(ii) the columns of H are linearly independent,

(iii) the number of rows of G plus the number of columns of H equals the number of columns of G which equals the number of rows of H,

and

(iv) $GH = 0$.

5. Show that there is no perfect 2-error correcting code of length 90 and size 2^{78} over $\{0,1\}$.

[*Hint:* Assume that the zero vector is a codeword. Consider the 88 words of weight 3 with 1 in the first two places, and show that each must be distant at most 2 from a codeword.]

6. For a word $x = x_1 \ldots x_N \in \{0,1\}^N$ define the *weight* $w(x)$ as the number of non-zero digits : $w(x) = \sharp\{i : x_i \neq 0\}$. For a linear (N,k) code $\mathcal{X}$ let A_i be the number of words in $\mathcal{X}$ of weight i $(0 \leq i \leq N)$. Define the *weight enumerator polynomial* $\mathbf{A}(\mathcal{X}, z) = \sum_0^N A_i z^i$.

Show that if you use $\mathcal{X}$ on a binary symmetric channel with error probability p, the probability of failing to detect an incorrect word is $(1-p)^N \left(\mathbf{A}\left(\mathcal{X}, \frac{p}{1-p}\right) - 1\right)$.

Show that the weight enumerator polynomials $\mathbf{A}(\mathcal{X}, z)$ and $\mathbf{A}(\mathcal{X}^+, z)$ of linear code $\mathcal{X}$ and its parity–check extention $\mathcal{X}$ (see Q. 2(b)) are related by $\mathbf{A}(\mathcal{X}^+, z) = 1/2 \left[(1+z)\mathbf{A}(\mathcal{X}, z) + (1-z)\mathbf{A}(\mathcal{X}, -z)\right]$.

Can you identify a code $\mathcal{X}_{\mathrm{ev}}$ with $\mathbf{A}(\mathcal{X}_{\mathrm{ev}}, z) = 1/2 \left[\mathbf{A}(\mathcal{X}, z) + \mathbf{A}(\mathcal{X}, -z)\right]$?

7. Prove that a 2-error correcting code of length 10 can have at most 12 codewords.

[*Hint:* The distance of the code must be ≥ 5. Suppose that it contains r codewords and extend it to an $[11, r]$ code of distance 6. List all codewords of the extended code as rows of an $r \times 11$ matrix. If column i in this matrix contains s_i zero's and $r - s_i$ one's then $6(r-1)r \leq \sum_{x \in \mathcal{X}^+} \sum_{x' \in \mathcal{X}^+} d(x, x') \leq 2 \sum_{i=1}^{11} s_i(r - s_i)$. The RHS is $\leq (1/2) \cdot 11\, r^2$ if r is even and $\leq (1/2) \cdot 11\, (r^2 - 1)$ if r is odd.]

8. Prove that there are 129 non-equivalent cyclic codes of length 128 (including the trivial codes, $\{0 \ldots 0\}$ and $\{0,1\}^{128}$). [*Hint:* Prove that $1 + X^{2^k} = (1+X)^{2^k}$.]

Find all cyclic codes of length 7.

If $\mathcal{X}$ is a cyclic code prove that the dual code is also cyclic and find its ıerator.

Lent 1996 YMS

COMMUNICATION THEORY: Example Sheet 1(O): Optional Extras

The Example sheets 1(O), 2(O) and 3(O) contain further examples to the course. They are intended to help deepening the knowledge of the course and provide a base for intensive revisions. Students may also try to prepare some of the optional examples for their supervisions in Lent Term. I intend to give, in due course, model solutions to most optional examples.

Some of questions are from previous years' Tripos papers: their style is essentially preserved, and they may require additional material which is not in the present course schedules.

Information sources and coding

1. Suppose that in the noiseless coding $I_m \to J_a$ one of the characters in J_a is always used as a word-space, so that the codewords consist of the m shortest distinct sequences of characters from J_{a-1} followed by the space character. Such a coding is decipherable. Verify that it satisfies the Kraft inequality.

2. Suppose the letters of a text with alphabet $I = \{1, \ldots, m\}$, which appear with frequencies $p_1, \ldots, p_m$, are to be coded into sequences of symbols from another alphabet, J, consisting of a symbols, $a < m$.

(a) Deduce upper and lower bounds for the minimal expected length of a decipherable coding.

(b) How would you construct a code with performance not worse than this upper bound?

(c) What is the condition on the p_i that this lower bound should be attained?

[**Hint**: *Use the Kraft inequality, or Shannon's NC theorem.*]

3. In the previous example, let $s_1, \ldots, s_m$ be lengths of the words that encode the corresponding letters.

(a) A subset B of $\{1, \ldots, m\}$ is given, of $2 \le k \le m$ elements. Prove that a decipherable coding such that

$$s_i = r \qquad (i \in B),$$

where r is the least integer with $ka^{-r} < 1$, minimizes $\max_{i \in B} s_i$ over all decipherable codings.

(b) Let frequencies $p_1, \ldots, p_m$ be ordered so that $p_1 \ge p_2 \ge \ldots \ge p_m > 0$, and suppose that there exists k such that $p_1 + \cdots + p_k = 1/2$. Let S be the random codeword length of a coding $I \to J^*$, and define the *median* of S to be the least integer s such that $\mathsf{P}(S \le s) \ge 1/2$. Prove that a coding as in (a), with $B = \{1, \ldots, k\}$, minimizes the median of S over all decipherable codings.

4. In examples 2 and 3, it is desired to find a decipherable code that minimizes the expected value of a^S. Establish the lower bound $\mathsf{E}\, a^S \ge \left(\sum_i \sqrt{p_i}\right)^2$ and characterize when equality occurs.

[**Hint:** *Employ the Cauchy-Schwartz inequality:*

$$\left|\sum_i x_i y_i\right| \le \left(\sum_i x_i^2\right)^{1/2} \left(\sum_i y_i^2\right)^{1/2},$$

with equality iff $x_i = cy_i$ for all i.]

Prove that an optimal code for the above criterion must satisfy $\mathsf{E}\, a^S < a\left(\sum_i \sqrt{p_i}\right)^2$.

5. a) Let $Z = X + Y$. Show that $h(Z|X) = h(Y|X)$. If X and Y are independent, prove that $\max[h(X), h(Y)] \le h(Z)$. [That is, adding an independent random variable adds uncertainty.] When is $h(Z) = h(X) + h(Y)$?

b) Let Z be defined as a disjoint *mixture* of X and Y. That is, X takes values in in set $\{1, \ldots, m\}$ with probabilities $p_X(u)$ and Y in $\{m+1, \ldots, m+m'\}$ with probabilities $p_Y(u)$, and

$$\begin{aligned} Z &= X, \text{ with probability } \alpha, \\ &= Y, \text{ with probability } 1 - \alpha. \end{aligned}$$

Find $h(Z)$ in terms of α, $h(X)$ and $h(Y)$. Show that $2^{h(Z)} \le 2^{h(X)} + 2^{h(Y)}$.

6. Suppose that a gastric infection is known to originate in exactly one of the m Cambridge restaurants, the probability it originates in the j^{th} being p_j. One has samples from all restaurants; by testing the pooled samples from a set of restaurants A one can determine with certainty whether the infection originates in A or its complement. Let $F(p_1, \ldots, p_m)$ denote the minimal expected number of such tests needed to locate the infection.

Show that

$$F \ge -\sum_j p_j \log_2 p_j$$

and determine the conditions under which the bound can be obtained.

7. a) Living near a sports ground, a student is able to detect, during a football match, the total number of goals scored but he (she) is not able to determine the score. The score will be announced in a television news programme soon after the match. Before the game starts, the student tries to evaluate the expected amount of information about the score which he will gain by the end of the match and that provided him by the television programme afterwards. It is known that both teams are equally able, and the a priori distribution of the total goal number is geometric, with parameter q. What are the answers?

[**Hint:** *After a few calculations, the student discovers that the answer to the first question is not changed if the teams are not equally able, but the distribution of total goals scored remains the same.*]

b) Give the answers when you have no information about the chances of the teams, but keep the above assumption about the total goal number distribution.

8. A helicopter commando crew has to install a surveillance unit on a road in enemy territory to observe vehicle movement during the night before a battle. A unit must be

installed at 8:00 pm and be taken back at 6:00 the next morning. There are two types of unit available: one which can register only the number of vehicles that have passed, and another which in addition records the hour and the minute each vehicle passes. But the second type of unit may be captured by the enemy with probability 1/2 whereas the first one is practically invisible. The decision as to which kind of unit should be used is taken on the basis of the expected amount of information gained next morning. The probability that a vehicle passes at a given one-minute interval is estimated as 1/4, and the probability that two or more of them pass at that interval is negligible. Assume that passing at standard non-overlapping one-minute intervals occurs independently. What is your suggestion? Explain why.

[**Note**: *It doesn't help you much to place the two units simultaneously: while capturing the more sophisticated unit, the enemy will almost certainly destroy the more simple one too, by bulldozing.*]

9. Prove that the entropy is a concave function of the probability-distribution vector: if $\mathbf{p}$ and $\mathbf{q}$ are two probability distributions and $\lambda \in [0,1]$ then $h(\lambda\mathbf{p}+(1-\lambda)\mathbf{q}) \geq \lambda h(\mathbf{p}) + (1-\lambda)h(\mathbf{q})$ and specify the cases of equality.

[**Hint:** *Consider a random variable θ taking values 0 and 1 with probabilities λ and $1-\lambda$ and connect θ with random variables X_0 and X_1 having distributions $\mathbf{p}$ and $\mathbf{q}$, respectively. Cf. Ex. 5 b). Then use a conditional entropy inequality.*]

The definition of the entropy raised a question of what properties one should require from a function $f(p_1,\dots,p_m)$ of a probability distribution in order to have $f = h$. A group of examples below address various aspects of this question. Function f is always asumed to be symmetric in its arguments.

10A*. Prove that if f possesses the following properties:

a) $F(m) = f(1/m,\dots,1/m)$ is monotone increasing in m,

b) if p_1, ..., p_m and q_1, ..., q_M are two probability distributions then

$$f(p_1q_1, p_1q_2, \dots, p_1q_M, p_2, p_3, \dots, p_m) = f(p_1,\dots,p_m) + p_1 f(q_1,\dots,q_M),$$

c) f is a continuous function of p_i when $p_i \geq 0$,

d) $f(1/2,1/2) = 1$,

then $f = h$.

10B*. Prove that if f possesses the following properties:

a) $f(p_1,p_2,p_3,\dots,p_m) = f(p_1+p_2,p_3,\dots,p_m) + (p_1+p_2) f\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\right)$,

b) $f(1/2,1/2) = 1$,

c) $f(p,1-p)$ is a continuous function of $p \in [0,1]$,

then $f = h$.

11*. It turns out that 'more homogeneous' probability distributions have a greater entropy. If $\mathbf{p} = (p_1,\dots,p_m)$ and $\mathbf{q} = (q_1,\dots,q_m)$ are two distributions then $\mathbf{p}$ is called more homogeneous than $\mathbf{q}$ ($\mathbf{p} \succ \mathbf{q}$) if, after rearranging the values in decreasing order:

$$p_1 \geq p_2 \geq \dots \geq p_m, \quad q_1 \geq q_2 \geq \dots \geq q_m,$$

one has $\sum_{i=1}^{k} p_i \leq \sum_{i=1}^{k} q_i$ for each $k = 1$, ..., m (and $\mathbf{q} \neq \mathbf{q}$). Show that if $\mathbf{p} \succ \mathbf{q}$ then $h(\mathbf{p}) > h(\mathbf{q})$.

12. A random variable U takes values u from a finite set I with probabilities $p(u)$. Prove that for any $x \in (0,1)$ the probability that $p(U) \leq x$ is bounded by $\dfrac{h(U)}{\log\ 1/x}$.

13. By using properties of the entropy prove that $\ln x \geq 1 - 1/x$, $x \geq 0$.

Although the entropy was introduced in the course only for random variables taking finite set of values, the definition is easy to extend to the case of a countably many values or a random variable with density f:

$$H = -\sum_j p(j) \log\ p(j), \quad H = -\int dx f(x) \log\ f(x);$$

the only condition is that the series and integral converge. The value of the integral is called the differential entropy of the corresponding random variable X and denoted $h_{\text{diff}}(X)$.

14A. Show that the geomtertic distribution on $\mathbf{Z}_+$ (the non-negative integers) has maximum entropy amongst all distribitions on $\mathbf{Z}_+$ with the same mean. [A random variable ν has a geometric distribution if $\mathbf{P}(\nu = j) = \rho(1-\rho)^j$, $j \in \mathbf{Z}_+$, where $0 < \rho < 1$.]

B*. Suppose that two non-negative random variables X and Y are related by $Y = X + \nu$, where ν is independent of X and is geometrically distributed on $\mathbf{Z}_+$. What distribution maximises the mutual entropy of X and Y under the constraint $\mathbf{E}X \leq K$? Show that this distribution can be realised by assigning to X value zero with a certain probability and letting it follow a geometric distribution (having an appropriate expectation) with the complimentary probability.

15. Evaluate $h_{\text{diff}}(X)$ in the following cases:

a) the exponential density $f(x) = \lambda \exp\ (-\lambda x) 1(x \geq 0)$.

b) the normal density $f(x) = (2\pi\sigma^2)^{-1/2} \exp\left(-\dfrac{(x-a)^2}{2\sigma^2}\right)$.

Show that the exponential density maximises h_{diff} among the probability densities on $[0,\infty)$ with a given mean and the normal density among the probability densities on $\mathbf{R}$ with a given variance.

16. An ordinary deck of cards containing 26 red and 26 black is shuffled and dealt out one card at a time without replacement. Let X_i be the colour of the i-th card.

a) Determine $h(X_1)$.

b) Determine $h(X_2)$.

c) Does $h(x_k|X_1,\dots,X_{k-1})$ increase or decrease?

d) Determine $h(X_1,\dots,X_{52})$.

17. Let U_0, U_1, ... be a Markov chain. Prove that $h(U_n|U_0)$ and $h(U_0|U_n)$ are monotone in n.

18. Suppose that U_t, $t = 0, \pm1, \pm2$, can take values 0 or 1, and that the probability that $U_t = 1$, conditional $U_{t-1}, U_{t-2}, \dots$ is b_j, where j is the time which has elapsed since

U last equalled 1. Assume that $0 < b_j < 1$ for $0 < j \leq N$, and $b_j = 1$ for $j > N$. Show that the sequence X_t, where X_t is the time which has elapsed since U last equalled 1, is a Markov chain, and calculate its information rate.

19. A binary source emits digits 0 or 1 according to the rule

$$P(X_t = k | X_{t-1} = j, X_{t-2} = i) = q_r,$$

where k, j, i and r take values 0 or 1, $r = k - j - i \pmod 2$, and $q_0 + q_1 = 1$. Determine the information rate of the source.

Also derive the information rate of a binary Bernoulli source, emitting digits 0 and 1 with probabilities q_0 and q_1.

Explain the relationship between your two results.

20. At each time unit a device reads the current version of a string of N characters each of which may be either 0 or 1. It then transmits the number of characters which are equal to 1. Between each reading the string is perturbed by changing one of the characters at random (from 0 or 1 or vice versa, with each character being equally likely to be changed). Determine an expression for the information rate of this source.

21. Suppose that the sequence X_t, $t = 0, 1, \ldots$ is a Markov chain. For any random variables Y_1, Y_2, Y_3 define

$$i(Y_1, Y_2 | Y_3) = h(Y_1 | Y_3) + h(Y_2 | Y_3) - h(Y_1, Y_2 | Y_3)$$

and

$$i(Y_1, Y_2) = h(Y_1) + h(Y_2) - h(Y_1, Y_2).$$

Show that

$$i(X_{t-1}, X_{t+1} | X_t) = 0 \text{ and hence } i(X_{t-1}, X_{t+1}) \leq i(X_t, X_{t+1}).$$

Show also that $i(X_t, X_{t+s})$ is non-increasing in s, for $s = 0, 1, 2, \ldots$.

22. The sequence $\{X_t : t = \ldots, -2, -1, 0, 1, 2, \ldots\}$ is the output of a Markov chain with a finite state space and unique invariant distribution. Let $h(Z|Y)$ denote, for random variables Z and Y, the conditional entropy of Z given Y. Quoting standard properties of entropy, show that

$$h(X_t | X_{t-1}) \leq h(X_t | X_{t-2}) \leq 2h(X_t | X_{t-1}).$$

The output of the source is modified by malfunction of the recording equipment: every third symbol is replaced by an unreadable splodge, *. Recording is started at a random time, so that

$$P(X_1 = *) = P(X_2 = *) = P(X_3 = *) = \tfrac{1}{3}.$$

Show that the information rate of the modified source is

$$\tfrac{1}{3}\{h(X_t | X_{t-1}) + h(X_t | X_{t-2})\}$$

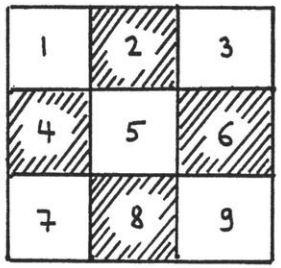
and deduce that not more than one-third of the information is lost. When is precisely $\frac{1}{3}$ lost?

For the case of a two-state Markov source with transition matrix

$$\Pi = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

find expressions for the information rate of the source, and of the source modified by illegibility of every third letter.

23. Find the information rate of a source represented by the Markov chain associated with a random walk of a king on the 3×3 chessboard:



Find the information rate for a rook, bishop (both kinds) and queen.

24. A stationary source emits symbols $0, 1, \ldots, m$ ($m \geq 4$ is an even number), according to a Markovian rule, with the following transition probabilities $p_{jk} = P(U_{n+1} = k \mid U_n = j)$

$$p_{jj+2} = 1/3,\ 0 \leq j \leq m-2,\ \ p_{jj+2} = 1/3,\ 2 \leq j \leq m,$$

$$p_{jj} = 1/3,\ 2 \leq j \leq m-2,\ \ p_{00} = p_{11} = p_{m-1m-1} = p_{mm} = 2/3.$$

The distribution of the first symbol is equiprobable. Find the information rate of the source. Does the result contradict Shannon's First Coding Theorem?

How does the answer change if m is odd? Can you use, for m odd, Shannon's First Coding Theorem to derive the information rate of the above source?

Lent 1996 YMS
E-mail: yms@statslab.cam.ac.uk

COMMUNICATION THEORY: Example Sheet 2(O): Optional Extras

The Example sheets 1(O), 2(O) and 3(O) contain further examples to the course. They are intended to help deepening the knowledge of the course and provide a base for intensive revisions. Students may also try to prepare some of the optional examples for their supervisions in Lent Term. I intend to give, in due course, model solutions to most optional examples.

Some of questions are from previous years' Tripos papers: their style is essentially preserved, and they may require additional material which is not in the present course schedules.

Information sources and coding

Although the lectures concentrate mainly on memoryless binary channels, many facts remain true for memoryless channels with more than two symbols, and even with non-equal numbers of symbols at the input and output (an example of which is occurrence of an intelligible symbol at the output). In particular, formula $C = \sup\, i(X,Y)$ holds where the supremum is over all possible input symbol distributiuons p_X. In examples 1–5 below you do not need to assume that the channel is binary.

1. Show that the capacity of a memoryless channel is always achieved by some input symbol distribution p_X.

[**Hint**: $i(X,Y)$ is a continuous function of p_X.]

2. Show that $h(Y)$ is a convex and $h(Y|X)$ a linear function of the input symbol probabilities $p_j = p(X_i = x_j)$.

3. A memoryless channel is called lossless if $h(Y|X) = 0$, deterministic if $h(Y|X) = 0$ and useless, if $h(X|Y) = h(X)$, whatever p_X. A channel is called noiseless if it is both lossless and deterministic. Show that the capacity of a lossless channel equals $\log\ m$ where m is the size of the input alphabet. Show that the capacity of the deterministic channel equals $\log\ s$ where s the size of the output alphabet. Show that the capacity of a useless channel eqials zero. Determine the input distributions that achieve capacity for all types of channels.

4. Prove that $0 \le C \le \min\,[\log\ m, \log s]$, where m and s are as in Question 3. In the case where $m = s$, prove the inverse assertion: $C = \log m$ only if the channel is lossless or deterministic (i.e., the cahnnel matrix is a permutation matrix). Prove that $C = 0$ only if the cahnnel is useless.

5. For a memoryless channel and the ideal observer decoding rule, show that the average probability of error is $\le \frac{1}{2} h(X|Y)$.

6. Consider the following binary channel *with memory*: $Y_j = X_j + Z_j \pmod 2$. Here X_j, Y_j, $Z_j = 0, 1$, and Z_1, Z_2, ..., is a sequence of random variables with

$$\mathsf{P}\big(Z_1 = Z_2 = \ldots = 0\big) = 1 - p,\ \ \mathsf{P}\big(Z_1 = Z_2 = \ldots = 1\big) = p,$$

where $0 < p < 1$. [I.e., Z_1, Z_2, ... is a repetitive Markov chain.] Show that the capacity of this channel is 1. Compare with the m.b.s.c. where the Z's are i.i.d. with $\mathsf{P}(Z_j = 0) = 1 - p$, $\mathsf{P}(Z_j = 1) = p$. Calculate the capacity of the channel when the Z's form an alternating Markov chain:

$$\mathsf{P}\big(Z_1 = 0,\ Z_{j+1} \ne Z_j,\ j \ge 1\big) = 1 - p,\ \ \mathsf{P}\big(Z_1 = 1,\ Z_{j+1} \ne z_j,\ j \ge 1\big) = p,$$

Questions 7–9 below are related to a *joint* source–channel coding theorem that puts together the assertions of the First and Second SCT's.

7. (The direct part) Suppose that a text emitted by a source with the AEP and information rate H is to be transmitted through a channel of capacity C. Prove that if $H + C < 1$ then there exists a sequence $\big(f_n, \widehat{f}_N\big)$ of codes f_n: $I^n \to \mathcal{X}_N \subset \{0,1\}^N$ and decoding rules $\widehat{f}_N$: $\{0,1\}^N \to I^n$, with $\lim\limits_{n\to\infty} \mathsf{P}_{\rm err} = 0$, where

$$\mathsf{P}_{\rm err} = \sum_{u^{(n)} \in I^n} \mathsf{P}_{\rm source}\big(U^{(n)} = u^{(n)}\big) \sum_{y^{(N)} \in \{0,1\}^N} \mathsf{P}_{\rm channel}\left(y^{(N)} \middle| f_N(u^{(n)})\right)$$

$$\times \mathbf{1}\left(\widehat{f}_N(y^{(N)} \ne u^{(n)}\right).$$

8*. (The converse part) For the source and channel as in Question 7, prove that if $H + C > 1$ then for any sequence $\big(f_n, \widehat{f}_N\big)$ of coding and decoding rules, $\liminf\limits_{n\to\infty} \mathsf{P}_{\rm err} > 0$.

[**Hint**: A useful exrecise is to introduce, for an arbitrary stationary source U_1, U_2, ..., the quantitites $J_n = n^{-1} h(U^{(n)})$ and $K_n = h(U_{n+1}|U^{(n)})$ and prove that both J_n and K_n are non-increasing with n and have, as $n \to \infty$, a common limit value (called the entropy of the source). For a source with the AEP, this value equals H, the information rate.]

9. A text is produced by a Bernoulli source with alphabet $I = \{1, 2, \ldots, m\}$ and probabilities p_1, p_2, ... p_m. It is desired to send this text reliably through a memryless binary symmetric channel with the tow error probability p^*. Write an essay explaining why reliable transmission is possible if

$$h(p_1, p_2, \ldots, p_m) + h(p^*, 1 - p^*) < 1,$$

and is impossible if

$$h(p_1, p_2, \ldots, p_m) + h(p^*, 1 - p^*) > 1.$$

10. Consider a memoryless channel with alphabet $\{0,1,2,3,4,5\}$ and the transition probabilities

$$p(y|x) = 1/2, \text{ if } y = x \pm 1 \bmod 5,$$
$$= 0, \text{ otherwise.}$$

Calculate the capacity of the channel.

The *zero-error* capacity of a channel is defined as the number of bits per channel symbol (or per channel use) that can be transmitted with error–probability zero. By transmitting 0 or 1 with probability 1/2 check that the zero–error capacity of the above channel is ≥ 1 and by considering codes of length 2 show that it is > 1.

[The precise value of the zero–error capacity here is $\frac{1}{2}\log 5$. This has been proven as recently as in 1979, after more than 20 years of efforts.]