Information Theory

Problem Sheet 1

(Most questions are from Cover & Thomas, the corresponding question numbers (as in 1st ed.) are given in brackets at the start of the question)

Notation: x, $\mathbf{x}$, $\mathbf{X}$ are scalar, vector and matrix random variables respectively.

The following expressions may be useful: $\sum_{n=1}^{\infty} r^n = \frac{r}{1-r}$ $\quad \sum_{n=1}^{\infty} nr^n = \frac{r}{(1-r)^2}$

1. [2.1] A fair coin is flipped until the first head occurs. Let x denote the number of flips required.
   (a) Find the entropy $H(x)$ in bits.
   (b) A random variable x is drawn according to this distribution. Find an "efficient" sequence of yes-no questions of the form "Is x contained in the set S?". Compare $H(x)$ to the expected number of questions required to determine x.

2. [~2.2] x is a random variable taking integer values. What can you say about the relationship between $H(x)$ and $H(y)$ if
   (a) $y = x^2$
   (b) $y = x^3$

3. [2.3] If $\mathbf{p}$ is an n-dimensional probability vector, what is the maximum and the minimum value of $H(\mathbf{p})$. Find all vectors $\mathbf{p}$ for which $H(\mathbf{p})$ achieves its maximum or minimum value.

4. We write $H(p)$ (with a scalar p) to denote the entropy of the Bernoulli random variable with probability mass vector $\mathbf{p} = [1-p \quad p]$. Prove the following properties of this function:
   (a) $H'(p) = \log(1-p) - \log p$
   (b) $H''(p) = \frac{-\log e}{p(1-p)}$
   (c) $H(p) \ge 2\min(p, 1-p)$
   (d) $H(p) \ge 1 - 4(p - ½)^2$
   (e) $H(p) \le 1 - 2\log e(p - ½)^2$

5. [2.5] Let x be a discrete random variable and $g(x)$ a deterministic function of it. Show that $H(g(x)) \le H(x)$ by justifying the following steps:

$$H(x, g(x)) \overset{(a)}{=} H(x) + H(g(x) \mid x) \overset{(b)}{=} H(x)$$
$$H(x, g(x)) \overset{(c)}{=} H(g(x)) + H(x \mid g(x)) \overset{(d)}{\ge} H(g(x))$$

6. [2.6] Show that if $H(y \mid x) = 0$, then y is a function of x, that is for all x with $p(x)>0$, there is only one possible value of y with $p(x,y) > 0$.

7. [~2.7] x_i is a sequence of i.i.d. Bernoulli random variables with $p(x_i = 1) = p$ where p is unknown. We want to find a function f that converts n samples of x into a smaller number, K, of i.i.d. Bernoulli random variables, z_i, with $p(z_i=1)=½$. Thus $z_{1:K}=f(x_{1:n})$ where K can depend on the values x_i.

   (a) Show that the following mapping for n=4 satisfies the requirements and find the expected value of K, $E(K)$.

   0000,1111→ignore; 1010→0; 0101→1; 0001,0011,0111→00;

   0010,0110,1110→01; 0100,1100,1101→10; 1000,1001,1011→11

   (b) Justify the steps in the following bound on $E(K)$

$$nH(p) \overset{(a)}{=} H(x_{1:n}) \overset{(b)}{\geq} H(z_{1:K}, K) \overset{(c)}{=} H(K) + H(z_{1:K} \mid K)$$
$$\overset{(d)}{=} H(K) + E\,K \overset{(e)}{\geq} E\,K$$

8. [2.10] Give examples of joint random variables x, y and z such that:

   (a) $I(x;y \mid z) < I(x;y)$

   (b) $I(x;y \mid z) > I(x;y)$

9. [2.12] We can define the "mutual information" between three variables as

$$I(x;y;z) = I(x;y) - I(x;y \mid z)$$

   (a) Prove that

$$I(x;y;z) = H(x,y,z) - H(x,y) - H(y,z) - H(z,x)$$
$$+ H(x) + H(y) + H(z)$$

   (b) Give an example where $I(x;y;z)$ is negative. This lack of positivity means that it does not have the intuitive properties of an "information" measure which is why I put "mutual information" in quotes above.

10. [2.17] Show that $\log_e(x) \geq 1 - x^{-1}$ for x>0.

11. [~2.16] x and y are correlated binary random variables with $p(x=y=0)=0$ and all other joint probabilities equal to 1/3. Calculate $H(x)$, $H(y)$, $H(x|y)$, $H(y|x)$, $H(x,y)$, $I(x;y)$.

12. [~2.22] If $x \to y \to z$ form a markov chain, and for y, the alphabet size $|\mathrm{Y}| = k$, show that $I(x;z) \leq \log k$. What does this tell you if $k = 1$?

13. [2.29] Prove the following and find the conditions for equality:

   (a) $H(x,y \mid z) \geq H(x \mid z)$

   (b) $I(x,y;z) \geq I(x;z)$

   (c) $H(x,y,z) - H(x,y) \leq H(x,z) - H(x)$

   (d) $I(x;z \mid y) \geq I(z;y \mid x) - I(z;y) + I(x;z)$

Information Theory

Problem Sheet 2

Notation: x, $\mathbf{x}$, $\mathbf{X}$ are scalar, vector and matrix random variables respectively.

1. Use the Kraft inequality to show that it is possible to construct a 4-ary instantaneous code with lengths {1, 1, 2, 2, 2, 2, 2, 2, 2}. Construct such a code for symbols that take the values A, B, …, H, I with probabilities {.15, .15, .1, .1, .1, .1, .1, .1, .1}.

   Calculate the entropy of the input symbols and the expected length of the codewords.

2. The four symbols A, B, C, D are encoded using the following sets of codewords. In each case state whether the code is (i) non-singular, (ii) uniquely decodable and (iii) instantaneous code.

   (a) {1, 01, 000, 001}

   (b) {0, 10, 000, 100}

   (c) {01, 01, 110, 100}

   (d) {0, 01, 011, 0111}

   (e) {10, 10, 0010, 0111}

3. In a complete D-ary code tree, the end of each branch is either a leaf node of else has D sub-branches. Show that the total number of leaf nodes is one more than a multiple of D–1.

4. For each value of D given below, find a D-ary Huffman code for the probability vector {0.25, 0.2, 0.15, 0.1, 0.1, 0.1, 0.1}. In each case calculate the expected code length. [Note that, to ensure a full code tree, you should add zero-probability symbols so that the total number of symbols is one more than a multiple of D–1]

   (a) D = 3

   (b) D = 4

   (c) D = 5

5. Find a binary instantaneous code with codeword lengths {2, 2, 3, 3, 3, 4, 4}. Find a probability mass vector for which the expected length of this code is equal to the entropy of the source.

6. [4.2] If $\{x_i\}$ is a stationary stochastic process, show that

$$H(x_i \mid x_{i-1}, x_{i-2}, \cdots, x_{i-n}) = H(x_i \mid x_{i+1}, x_{i+2}, \cdots, x_{i+n}) .$$

   In other words, the conditional entropy given the previous n samples is the same as the conditional entropy given the next n samples.

7. (a) [4.5] Determine the stationary distribution and the entropy rate, $H(\mathsf{X})$, of a Markov process with two states, 0 and 1 and transition matrix

$$T = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix} .$$

   (b) Find the values of p and q that maximize $H(\mathsf{X})$.

   (c) If q=1,

   (i) find the maximum value of $H(\mathsf{X})$ and the value of p that attains it.

   (ii) define $s(n)$ to be the number of sequences of length n with non-zero probability. If $s_0(n)$ and $s_1(n)$ are the number of sequences starting with 0 and 1 respectively, show that $s_1(n) = s_0(n-1) = s(n-2)$. Hence show that $s(n) = s(n-1) + s(n-2)$ and derive an explicit expression for $s(n)$.

   (iii) explain why we must have $H(\mathsf{X}) \le \lim_{n\to\infty}\left(n^{-1} \log s(n)\right)$ and calculate the value on the right hand side.

8. If $x_i \in \{A, B\}$ are i.i.d. with probability mass vector $\{0.9, 0.1\}$, determine $H(x_i)$.

   Using binary Huffman codes, determine the coding redundancy (i.e. the difference between the entropy and the number of bits used per symbol) when (a) each x_i is encoded individually, (b) pairs of x_i are coded together (i.e. x_1x_2 followed by x_3x_4 etc) and (c) triplets of x_i are coded together.

Information Theory

Problem Sheet 3

Notation: *x*, **x**, **X** are scalar, vector and matrix random variables respectively.

1. The Z channel. The Z-channel has binary input and output alphabets and transition probabilities $p(y|x)$ given by the following matrix:

$$Q = \begin{bmatrix} 1 & 0 \\ 1/2 & 1/2 \end{bmatrix} \quad x, y \in \{0,1\}$$

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

2. Calculate the capacity of the following channel with probability transition matrix

$$Q = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 0 & 1/2 & 1/2 \\ 1/2 & 0 & 1/2 \end{bmatrix} \quad x, y \in \{0,1,2\}$$

3. Differential entropy. Evaluate the differential entropy $h(X) = -\int f \ln f$ for the following:
(a) The exponential density $f(x) = \lambda e^{-\lambda x}$, $x \geq 0$.
(b) The sum of X_1 and X_2; where X_1 and X_2 are independent normal random variables with means μ_i and variances σ_i^2 for i = 1, 2.

4. Parallel channels and waterfilling. Consider a pair of parallel Gaussian channels, i.e.,

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} + \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}$$

where

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim N\left(0, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}\right)$$

and there is a power constraint $E(X_1^2 + X_2^2) \leq 2P$. Assume that $\sigma_1^2 \geq \sigma_2^2$. At what power does the channel stop behaving like a single channel with noise variance σ_2^2 and begin behaving like a pair of channels?

5. Rate-distortion function. Let

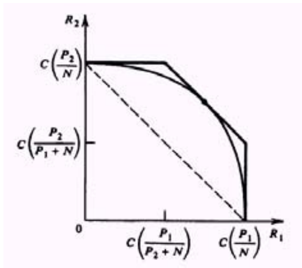
$$D(R) = \min_{p(\hat{x}|x): I(X;\hat{X}) \leq R} Ed(X, \hat{X})$$

be the rate-distortion function.
(a) Is $D(R)$ a decreasing or increasing function of R?
(b) Is $D(R)$ convex or concave in R?

6. Describe the capacity region of a two-user multiple access channel. Interpret the corner points (i.e., why can one of the users achieve the capacity as if the other user were absent?) Verify the following equality for the corner point:

$$C\left(\frac{P_1}{N}\right) + C\left(\frac{P_2}{P_1 + N}\right) = C\left(\frac{P_1 + P_2}{N}\right)$$

where $C(x) = (\log(1+x))/2$ is the capacity function.



7. Slepian-Wolf region for binary sources. Let X_i be i.i.d. Bernoulli(p). Let Z_i be i.i.d. Bernoulli(r), and let Z be independent of X. Finally, let Y = X⊕Z (mod 2 addition). Let X be described at rate R_1 and Y be described at rate R_2: What region of rates allows recovery of X, Y with probability of error tending to zero?

8. Describe the capacity region of a two-user Gaussian broadcast channel and sketch it.