IMPERIAL COLLEGE LONDON

EE4-40
EE9-CS7-26
EE9-SO20

DEPARTMENT OF ELECTRICAL AND ELECTRONIC ENGINEERING
EXAMINATIONS 2013

MSc and EEE/EIE PART IV: MEng and ACGI

Corrected Copy

INFORMATION THEORY

Tuesday, 14 May 10:00 am

Time allowed: 3:00 hours

There are FOUR questions on this paper.

Answer ALL questions.

All questions carry equal marks

Any special instructions for invigilators and information for candidates are on page 1.

Examiners responsible First Marker(s) : C. Ling

Second Marker(s) : D. Gunduz

© Imperial College London

Information for students

Notation:

(a) Random variables are shown in Tahoma font. X, $\mathbf{\mathsf{x}}$, $\mathbf{\mathsf{X}}$ denote a random scalar, vector and matrix respectively.

(b) The size of a set A is denoted by $|A|$.

(c) By default, the logarithm is to the base 2.

(d) $\oplus$ denotes the exclusive-or operation, or modulo-2 addition.

(e) "i.i.d." means "independent identically distributed".

(f) $C(x) = \frac{1}{2}\log_2(1+x)$ is the capacity function for the Gaussian channel in bits/channel use.

The Questions

1. Basics of information theory.

a) Given probability mass vectors **p** and **q**, both of length m,

i) Write down the entropy formula of $H(\mathbf{p})$. Given an interpretation of the entropy.

ii) What are the maximum and minimum of $H(\mathbf{p})$? When are they achieved?

iii) For $\mathbf{p} = [\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}]$ and $\mathbf{q} = [\frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{2}]$, compute the relative entropy $D(\mathbf{p}||\mathbf{q})$ and $D(\mathbf{q}||\mathbf{p})$.

iv) Give properties of a "distance" and based on iii) say if the relative entropy is a distance or not.

[12]

b) Let X be a random variable taking integer values. What can you say about the relationship between $H(X)$ and $H(Y)$ (justification is needed) if:

i) $Y = X^2$

ii) $Y = X^3$

[6]

c) If $X \to Y \to Z$ forms a Markov chain, and for Y, the alphabet size $|Y| = k$, show that $I(X; Z) \leq \log k$. What does this tell you if $k = 1$?

[7]

2. Source coding.

a) Rate-distortion of Gaussian sources. Assume $X \sim N(0,\sigma^2)$ and $E(X-\hat{X})^2 \le D$. Justify each step of the following derivations.

i) Lower bound on mutual information.

$$I(X;\hat{X}) \overset{(1)}{=} h(X) - h(X \mid \hat{X}) \overset{(2)}{=} \tfrac{1}{2}\log 2\pi e\sigma^2 - h(X-\hat{X} \mid \hat{X})$$

$$\overset{(3)}{\ge} \tfrac{1}{2}\log 2\pi e\sigma^2 - h(X-\hat{X}) \overset{(4)}{\ge} \tfrac{1}{2}\log 2\pi e\sigma^2 - \tfrac{1}{2}\log\left(2\pi e \operatorname{Var}\left(X-\hat{X}\right)\right)$$

$$\overset{(5)}{\ge} \tfrac{1}{2}\log 2\pi e\sigma^2 - \tfrac{1}{2}\log 2\pi e D$$

$$\overset{(6)}{\Rightarrow} I(X;\hat{X}) \ge \max\left(\tfrac{1}{2}\log\frac{\sigma^2}{D}, 0\right)$$

ii) Achievability. To show that we can find a distribution $p(\hat{x},x)$ that achieves the lower bound, we construct a test channel that introduces distortion $D<\sigma^2$ shown in Fig. 2.1.

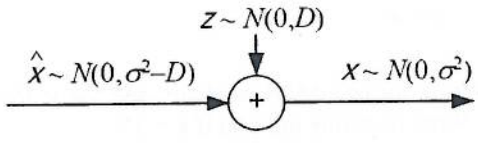


Fig. 2.1. Test channel.

$$I(X;\hat{X}) = h(X) - h(X \mid \hat{X}) = \tfrac{1}{2}\log 2\pi e\sigma^2 - h(X-\hat{X} \mid \hat{X})$$

$$\overset{(7)}{=} \tfrac{1}{2}\log 2\pi e\sigma^2 - h(Z \mid \hat{X}) \overset{(8)}{=} \tfrac{1}{2}\log\frac{\sigma^2}{D}$$

$$\overset{(9)}{\Rightarrow} I(X;\hat{X}) = \max\left(\tfrac{1}{2}\log\frac{\sigma^2}{D}, 0\right)$$

$$\overset{(10)}{\Rightarrow} D(R) = \frac{\sigma^2}{2^{2R}}$$

[10]

b) Shannon code. Given the probability mass vector $\mathbf{p} = [p_1, p_2, \dots, p_n]$ for an n-symbol source X, where $p_1 \ge p_2 \ge \cdots \ge p_n$ are sorted in decreasing order. Calculate the partial sum of the probabilities of the first $i-1$ symbols:

$$F_i = \sum_{j=1}^{i-1} p_j$$

Shannon's algorithm of source coding encodes the i-th symbol x_i into a codeword by rounding off the binary representation of $F_i \in [0,1]$ to $l_i = \lceil -\log_2(p_i) \rceil$ bits.

i) Show that the average codeword length L_S satisfies the following bound

$$H(X) \leq L_S \leq H(X) + 1$$

where $H(X)$ is the entropy of the source.

ii) Construct the code for the probability distribution $\mathbf{p} = [0.5, 0.25, 0.125, 0.125]$.

iii) Show that Shannon's code is an instantaneous code. Your answer should be general, i.e., not limited to the example in ii).

[15]

3. Channel coding.

a) Converse of the channel coding theorem. Consider the channel model shown in Fig. 3.1. Justify each step of the following proof.

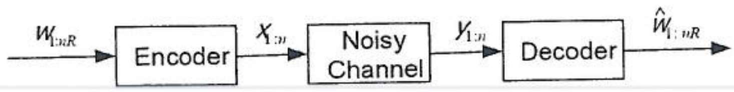


Fig. 3.1. Channel model.

$$
\begin{aligned}
nR &\overset{(1)}{=} H(\mathbf{w}) = H(\mathbf{w}\mid\mathbf{y}) + I(\mathbf{w};\mathbf{y}) \\
&\overset{(2)}{\le} H(\mathbf{w}\mid\mathbf{y}) + I(\mathbf{x}(\mathbf{w});\mathbf{y}) \\
&\overset{(3)}{\le} 1 + nRP_e^{(n)} + I(\mathbf{x};\mathbf{y}) \\
&\overset{(4)}{\le} 1 + nRP_e^{(n)} + nC \\
&\overset{(5)}{\Rightarrow} P_e^{(n)} \ge \frac{R - C - n^{-1}}{R} \underset{n\to\infty}{\to} 1 - \frac{C}{R} > 0 \quad \textit{if} \quad R > C
\end{aligned}
$$

$\overset{(6)}{\Rightarrow}$ For large (hence for all) n, $P_e^{(n)}$ has a lower bound of $(R–C)/R$ if w equiprobable.

[6]

b) Bandlimited channel. Consider a continuous channel bandlimited to bandwidth $[-W,W]$ and subject to white Gaussian noise with double-sided power spectral density $N_0/2$. Let T be the signal duration, in the sense that most energy of the signal falls into this time interval. Reproduce the Shannon capacity formula by following the steps given below:

i) Sample the signal at the Nyquist rate. Compute the energy constraint per sample E_s.

ii) Compute the noise power P_N associated with each sample.

iii) Compute the channel capacity C in bits/second.

[9]

c) Gaussian mutual information. Suppose X, Y, and Z are Gaussian random variables, each with zero mean and unit variance. Further, suppose that (X, Y, Z) are jointly Gaussian and that $X \to Y \to Z$ forms a Markov chain. Let X and Y have correlation coefficient ρ_1 and let Y and Z have correlation coefficient ρ_2.

i) Derive the correlation coefficient ρ_3 between X and Z.

ii) Find $I(X; Z)$.

Hint: Use the fact that $E[X\mid Y] = \rho_1 Y$ for jointly Gaussian random variables X, Y, each with zero mean and unit variance.

[10]

4. Network information theory.

a) Broadcast channel.

i) With the help of a diagram, explain what is a broadcast channel. Give an example of the broadcast channel.

ii) Consider the two-user scalar Gaussian broadcast channel. Write down the capacity region and describe the coding/decoding strategy.

[8]

b) Multi-access channel.

i) With the help of a diagram, explain what is a multi-access channel. Give an example.

ii) Write down the capacity region of the m-user multi-access channel with additive white Gaussian noise, where the users have equal powers P, and the noise power is N.

iii) Specialize to the two-user case. Draw its capacity region, and explain why CDMA is better than FDMA and TDMA.

[9]

c) Now a new user with power P_0 wishes to join the m-user multi-access channel.

i) At what rate R_0 can he send information without disturbing the other users?

ii) What should his power P_0 be so that the new user's rate is equal to the combined communication rate $C(mP/N)$ of all the other users?

[8]

B — Book work
A — Application
E — New example
T — New theory

ANSWERS

1\. a)

i) $H(p) = -\sum_{i=1}^{m} p(i) \log p(i)$ [3 B]

Entropy is the average Shannon information content, or the amount of uncertainty before we know the value.

ii) $\max H(p) = \log m$ iff all elements of p are equal.

$\min H(p) = 0$ iff only one element is nonzero. [2 A]

iii) $D(p\|q) = 0.35$

$D(q\|p) = 0.424$ [4 E]

iv) $D(p\|q) \neq D(q\|p)$ in general. Therefore, strictly speaking, $D(p\|q)$ is not a distance, although it can measure the similarity between p and q. [3 B]

b) $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$ [2 A]

but $H(Y|X) = 0$ since Y is a function of X.

So $H(Y) = H(X) - H(X|Y) \leq H(X)$ with equality iff $H(X|Y) = 0$. This is true only if X is a function of Y. This means Y has to be a one-to-one function of X.

i) $H(Y) \leq H(X)$ since $Y = X^2$ is not one-to-one. [2 A]

ii) $H(Y) = H(X)$ since $Y = X^3$ is one-to-one. [2 A]

c) By data processing theorem, $I(X;Z) \leq I(X;Y)$ [2 A]

$= H(Y) - H(Y|X) \leq H(Y) \leq \log k$ (uniform bound). [2 A]

If $k=1$, $\log k = 0$. So $I(X;Z) = 0$, which means X and Z are independent. [3 E]

2. a) i)

(1) definition of $I(X;\hat{X})$. [1 each, B]

(2) Gaussian entropy; Shift doesn't change entropy.

(3) Conditioning reduces entropy.

(4) Given variance, Gaussian has max entropy.

(5) $Var(X-\hat{X}) \le D$.

(6) Algebra and $I(X;\hat{X}) \ge 0$

ii) (7) $\hat{X} + Z = X$

(8) $h(Z|\hat{X}) = h(Z) = \frac{1}{2}\log 2\pi e D$

(9) The lower bound is achievable $\Rightarrow$ $\geq$ becomes $=$.

(10) definition of rate-distortion.

b) i) Since $l_i = \lceil -\log p_i \rceil$, we have [5T]

$$-\log p_i \le l_i \le -\log p_i + 1. \quad (*)$$

Average length

$$-\sum p_i \log p_i \le L_S = \sum p_i l_i \le -\sum p_i \log p_i + 1$$

$$H(S) \le L_S \le H(S) + 1.$$

ii) [5T]

Symbol	p_i	F_i	binary	l_i	codeword
1	0.5	0	0.0	1	0
2	0.25	0.5	0.10	2	10
3	0.125	0.75	0.110	3	110
4	0.125	0.875	0.111	3	111

iii) [difficult]. From (*), we have [5T]

$$2^{-l_i} \le p_i < 2^{-(l_i - 1)}.$$

Thus, F_j $(j > i)$ must be larger than F_i by more than 2^{-l_i}, therefore F_j $(j>i)$ must differ from F_i by one bit in the first l_i bits, in binary expansion. Thus their codewords also differ by at least one bit. This means no codewords are a prefix of another.

3.

a) [1 each B]

(1) input is uniform.

(2) $W \to X \to Y$ form a Markov chain.

(3) Fano's inequality.

(4) $I(X;Y) \le nC$

(5) algebra

(6) (6) (5) is true for large n. It is also true for small n, since we can concatenate several short blocks to form a long block.

b) i) By sampling at Nyquist rate, we obtain 2WT samples.

~~Let~~

$$E_s \cdot 2WT \le PT$$ [3 B]

$$\Rightarrow E_s \le \frac{P}{2W}$$

ii) Noise power $P_N = \frac{N_0}{2} \cdot 2W \cdot \frac{T}{2WT} = \frac{N_0}{2}$ [3 B]

iii)

$$C = \frac{1}{2}\log\left(1 + \frac{P/2W}{N_0/2}\right)\frac{2WT}{T}$$ [3 B]

$$= W\log\left(1 + \frac{P}{WN_0}\right)$$

c) i) $$\rho_3 = \frac{E[XZ]}{\sigma_x \sigma_z} = \frac{E\{E[XZ|Y]\}}{\sigma_x \sigma_z}$$ [5 T]

Obvious, but $\sigma_x, \sigma_y, \rho_z$ not defined.

$$= \frac{E\{E[X|Y]\,E[Z|Y]\}}{\sigma_x \sigma_z}$$ Markov chain

$$= \frac{E\{\rho_1 Y \cdot \rho_2 Y\}}{\sigma_x \sigma_z}$$

$$= \frac{\rho_1 \rho_2 \sigma_y^2}{\sigma_x \sigma_z}$$ $\sigma_x = \sigma_y = \sigma_z = 1$

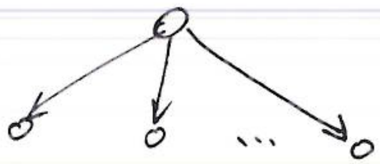
$$= \rho_1 \rho_2$$

ii) The covariance matrix

$$Cov(X, Z) = \begin{pmatrix} 1 & \rho_3 \\ \rho_3 & 1 \end{pmatrix}$$ [5 T]

$$\begin{aligned} I(X; Z) &= h(X) + h(Z) - h(X, Z) \\ &= \tfrac{1}{2}\log(2\pi e) + \tfrac{1}{2}\log(2\pi e) - \tfrac{1}{2}\log 2\pi e(1-\rho_3^2) \\ &= \tfrac{1}{2}\log\frac{2\pi e}{1-\rho_3^2}. \end{aligned}$$

4. a) i) Broadcast channel: [4 B]



One sender, many receivers. The sender can send different messages to different users.

Example: Mobile downlink.

ii) Capacity region: [4 B]

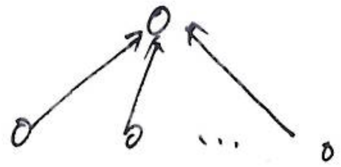
$$R_1 \le C\left(\frac{\alpha P}{N_1}\right) \qquad 0 \le \alpha \le 1$$

$$R_2 \le C\left(\frac{(1-\alpha)P}{\alpha P + N_2}\right)$$

Encoding: One codebook with power αP at rate R_1, another with power $(1-\alpha)P$ at rate R_2, send sum of two codewords.

Decoding: Bad receiver Y_2 treats Y_1 as noise; Good receiver Y_1 first decodes Y_2, subtracts it out, then decodes own message.

b) i) Multi-access channel



[3B]

One reciever, many senders. Example: CDMA uplink.

ii) Capacity region

$$R_i < c\left(\frac{P}{N}\right)$$

[3B]

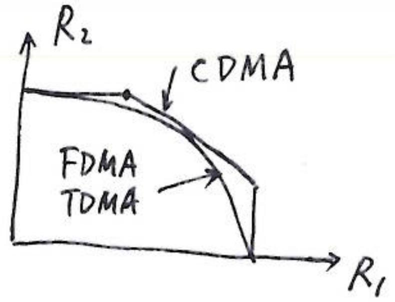
$$R_i + R_j < c\left(\frac{2P}{N}\right)$$

$$R_i + R_j + R_k < c\left(\frac{3P}{N}\right)$$

$$\vdots$$

$$\sum_{i=1}^{m} R_i \le c\left(\frac{mP}{N}\right)$$

iii) Two-user:



[3B]

c) i) If the new user can be decoded while treating other users as noise, then it is fine. Hence

$$R_0 < \frac{1}{2}\log\left(1+\frac{P_0}{mP+N}\right)$$

[4T]

ii)

$$\frac{1}{2}\log\left(1+\frac{P_0}{mP+N}\right) = \frac{1}{2}\log\left(1+\frac{mP}{N}\right)$$

[4T]

$$\frac{P_0}{mP+N} = \frac{mP}{N}$$

$$P_0 = (mP+N)\frac{mP}{N}$$