

# Info Theory Supervision 1

**Topics: Definition of information, Entropy (one RV), Weighing Problems, Unique Decodability, Huffman coding**

Key to questions types:

U: Core understanding (regurgitating notes)

S: Standard (need to engage brain but should not take long)

C: More challenging (errr... more challenging)

- 1 U For each of the properties Shannon identified for information (continuous, additive, symmetric), explain why it is necessary.
- 2 U Write a short argument for why the definition of Shannon Information is a meaningful representation of information, targeted at a fellow Part II student who is not taking Information Theory.
- 3 S Calculate the entropy in bits for each of the following random variables:
  - a) Pixel values in an image whose possible grey values are all the integers from 0 to 255 with uniform probability.
  - b) Humans classified according to whether they are, or are not, mammals.
  - c) Gender in a tri-sexed insect population whose three genders occur with probabilities  $1/4$ ,  $1/4$ , and  $1/2$
  - d) A population of persons classified by whether they are older, or not older, than the population's median age.
- 4 S We argued that Shannon Information should be both logarithmic and increase when the probability of an event decreases. Comment on whether the following could be valid definitions of information:
  - a)  $\log(1 - P(x))$
  - b)  $\log\left(\frac{1}{P(x)^2}\right)$
- 5 S Prove the maximum entropy under the constraint of an average value takes the exponential form stated in Lecture 1.
- 6 S Using Lagrangian Multipliers, maximise:
  - a)  $f(x, y) = (x - 1)^2 + (y - 5)^2$  under the constraint  $3x + 4y = 0$
  - b)  $f(x, y) = x^2 + y^2$  under the constraint  $x^2 + 4y^2 = 16$
  - c)  $f(x, y, z) = xyz$  under the constraints  $x^2 + y^2 + z^2 = 12$  **and**  $x = y$
- 7 S For the weighing problem with 10 balls, one of which is heavy, draw out a possible solution to finding the heavy ball in no more than 3 weighings. Now annotate each branch from each weighing with the information gained by following that branch. At each leaf, compute the total information gained for that branch.
- 8 S [Mackay ex 4.9] While some people, when they first encounter the weighing problem with 12

balls and the three-outcome balance, think that weighing six balls against six balls is a good first weighing, others say ‘no, weighing six against six conveys no information at all’. Explain to the second group why they are both right and wrong. Compute the information gained about which is the odd ball, and the information gained about which is the odd ball and whether it is heavy or light

- 9 C [Mackay 4.13]. Find a solution to the general N-ball weighing problem in which exactly one of N balls is odd. Show that in W weighings, an odd ball can be identified from among  $N = (3^W - 3)/2$  balls
- 10 S [Mackay 5.19] Is the code {00, 11, 0101, 111, 1010, 100100, 0110} uniquely decodable? If not, give an example of an ambiguous message.
- 11 S [Mackay 5.20] Is the ternary code {00, 012, 0110, 0112, 100, 201, 212, 22} uniquely decodable?
- 12 S [Mackay 5.21] ] Make Huffman codes for  $X^2$ ,  $X^3$  and  $X^4$  where  $A_X = 0, 1$  and  $P_X = 0.9, 0.1$ . Compute their expected lengths and compare them with the entropies  $H(X^2)$ ,  $H(X^3)$  and  $H(X^4)$ . Repeat this exercise for  $X^2$  and  $X^4$  where  $P_X = 0.6, 0.4$ .
- 13 C Is it possible for two people to use the same set of alphabet probabilities to produce Huffman codes with a differing set of lengths?
- 14 C Consider the following symbol encoding scheme:
  1. Each input symbol is associated with a probability of occurrence and an initially blank codeword.
  2. Partition the set of input symbols into two subsets, where the sum of the probabilities in each subset is as close to equal as possible.
  3. Append a 0 to every codeword in the first subset and a 1 to every codeword in the second subset
  4. Recurse on each subset until there is only one symbol left in each partition

Does this produce instantaneous codes? How does it compare to Huffman coding? (Note the full analysis of the average codeword length is non-trivial and not required. But that shouldn't stop you trying).

# Info Theory Supervision 2

**Topics: Source coding theorem, Arithmetic Coding, Lempel-Ziv, Error Correcting Codes**

Key to questions types:

U: Core understanding (regurgitating notes)

S: Standard (need to engage brain but should not take long)

C: More challenging (errr... more challenging)

- 1 U Prove the Kraft inequality, and that for any set of codes satisfying it there will be a uniquely decodable prefix code.
- 2 U Explain how Arithmetic coding can get closer to the entropy of a source than Huffman coding. Under what circumstances does this occur?
- 3 S Explain why a terminating character is included in Arithmetic coding. If it were not included, what information would you need to decode a message? If it were included, what effect does the probability assigned to it have on the code?
- 4 U What information must you convey to another party so that they can successfully decode a message you have encoded using Arithmetic coding?
- 5 S Compress the sequence “BAABBBBAAB” using (i) Huffman coding, (ii) arithmetic coding where the full message is known in advance, and (iii) Lempel-Ziv compression with a 3-bit dictionary. Comment on what you find.
- 6 P [Mackay ex. 6.3, pg 118]. Compare the following two techniques for generating random symbols from a nonuniform distribution  $\{p_0, p_1\} = \{0.99, 0.01\}$ :
  - a) The standard method: use a standard random number generator to generate an integer between 1 and  $2^{32}$ . Rescale the integer to  $(0, 1)$ . Test whether this uniformly distributed random variable is less than 0.99, and emit a 0 or 1 accordingly.
  - b) Arithmetic coding using the correct model, fed with standard random bits. Roughly how many random bits will each method use to generate a thousand samples from this sparse distribution?
- 7 P [Mackay ex 6.4. Pg 119]. ] Prove that *any* uniquely decodable code from  $\{0, 1\}^+$  to  $\{0, 1\}^+$  necessarily makes some strings longer if it makes some strings shorter.
- 8 P [Mackay ex 6.5. Pg 120]. Encode the string 000000000000100000000000 using the basic Lempel-Ziv algorithm.
- 9 P [Mackay ex 6.6. Pg 120]. Decode the string 00101011101100100100011010101000011 that was encoded using the basic Lempel-Ziv algorithm.

- 10 S What is the Hamming distance? Calculate the Hamming distance between the codewords 010101100011 111110001100. Indicate an application of the Hamming distance in error-correction.

# Info Theory Supervision 3

**Topics: Correlated RVs, Noisy coding, Gaussian channels, KL Divergence**

Key to questions types:

U: Core understanding (regurgitating notes)

S: Standard (need to engage brain but should not take long)

C: More challenging (errr... more challenging)

- 1 U Define joint entropy, conditional entropy and mutual information and show how they are related. Why is mutual information so important in Information Theory?
- 2 S Consider two independent integer-valued random variables,  $X$  and  $Y$ . Variable  $X$  takes on only the values of the eight integers  $\{1, 2, \dots, 8\}$  and does so with uniform probability. Variable  $Y$  may take the value of any positive integer  $k$ , with probabilities  $P\{Y = k\} = 2^{-k}$ ,  $k = 1, 2, 3, \dots$ 
  - a) Which random variable has greater uncertainty? Calculate both entropies  $H(X)$  and  $H(Y)$ .
  - b) What is the joint entropy  $H(X, Y)$  of these random variables, and what is their mutual information  $I(X; Y)$ ?
- 3 U In the proof of the noisy channel theorem, we created a code randomly and then used a 'typical set decoder' for it. In practice we would never do this so why do we do so in the proof, and why would the conclusions have any validity on a real code?
- 4 S There are three regions of interest with the noisy channel theorem: the first represents error-free communication rates, the second error-bounded communication rates and the third unachievable communication rates. Explain why the second region would ever be of interest.
- 5 S  $Y$  and  $Z$  are two continuous random variables.  $Y$  has an exponential probability density distribution  $p(x)$  over  $x \in [0, \infty]$ :  $p(x) = e^{-x}$ .  $Z$  has a uniform probability density distribution:  $p(x) = 1/\alpha$  for  $x \in [0, \alpha]$ , else  $p(x) = 0$ . Calculate the differential entropies  $h(Y)$  and  $h(Z)$  for these two continuous random variables, and find the value of  $\alpha$  for which these differential entropies are the same. Sketch these distributions.
- 6 C Consider a binary symmetric channel with error probability  $p$  that any bit may be flipped. Two possible error-correcting coding schemes are available: Hamming, or simple repetition.
  - a) Without any error-correcting coding scheme in place, state all the conditions that would maximise the channel capacity. Include conditions on the error probability  $p$  and also on the probability distribution of the binary source input symbols.
  - b) If a (7/4) Hamming code is used to deliver error correction for up to one flipped bit in any block of seven bits, provide an expression for the residual error probability  $P_e$  that such a scheme would fail.

- c) If repetition were used to try to achieve error correction by repeating every message an odd number of times  $N = 2m + 1$ , for some integer  $m$  followed by majority voting, provide an expression for the residual error probability  $P_e$  that the repetition scheme would fail.
- 7 U Contrast discrete and differential entropy. Prove that a Normal distribution maximises differential entropy for a given variance.
- 8 S Consider a noisy continuous communication channel of bandwidth  $W = 1$  MHz, which is perturbed by additive white Gaussian noise whose total spectral power is  $N_0W = 1$ . Continuous signals are transmitted across such a channel, with average transmitted power  $P = 1,000$ . Give a numerical estimate for the channel capacity, in bits per second, of this noisy channel. Then, for a channel having the same bandwidth  $W$  but whose signal-to-noise ratio  $P/N_0W$  is four times better, repeat your numerical estimate of capacity in bits per second.
- 9 S Consider a Gaussian channel. Are there any limits to the capacity of the channel if we increase
- a) the signal to noise ratio?
  - b) the bandwidth of the signal?
- 10 U Explain what is meant by the Kullback-Liebler divergence, and why it is a divergence and not a distance.
- 11 U What is meant by forward and reverse KLD in machine learning?