# E4.40/SO20 – Information Theory

**Problem Sheet** 1

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

**Notation:** We use a sans-serif font for random variables: $\mathsf{x}$, $\mathbf{\mathsf{x}}$, $\mathbf{\mathsf{X}}$ 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 $\mathsf{x}$ denote the number of flips required.

   (a) Find the entropy $H(\mathsf{x})$ in bits.

   (b) A random variable $\mathsf{x}$ is drawn according to this distribution. Find an "efficient" sequence of yes-no questions of the form "Is $\mathsf{x}$ contained in the set S?". Compare $H(\mathsf{x})$ to the expected number of questions required to determine $\mathsf{x}$.

2. [~2.2] $\mathsf{x}$ is a random variable taking integer values. What can you say about the relationship between $H(\mathsf{x})$ and $H(\mathsf{y})$ if

   (a) $\mathsf{y} = \mathsf{x}^2$

   (b) $\mathsf{y} = \mathsf{x}^3$

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 $\mathsf{x}$ be a discrete random variable and $g(\mathsf{x})$ a deterministic function of it. Show that $H(g(\mathsf{x})) \le H(\mathsf{x})$ by justifying the following steps:

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

6. [2.6] Show that if $H(\mathsf{y} \mid \mathsf{x}) = 0$, then $\mathsf{y}$ is a function of $\mathsf{x}$, 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 $E(K)$

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

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

$$\begin{aligned} I(x;y;z) = {} & H(x,y,z) - H(x,y) - H(y,z) - H(z,x) \\ & + H(x) + H(y) + H(z) \end{aligned}$$

(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 $|\mathcal{Y}| = k$, 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) $I(x;z \mid y) \geq I(z;y \mid x) - I(z;y) + I(x;z)$

# E4.40/SO20 – Information Theory

**Solution Sheet** 1

1. (a) $H(x)$=2

   (b) Ask if $x$ = 1, 2, 3, … in turn. Expected number of questions is 2.

2. $H(x,y)$=$H(x)$+$H(y|x)$=$H(Y)$+$H(x|y)$, but $H(y|x)$=0 since $Y$ is a function of $x$ so $H(y)$= $H(x)$ – $H(x|y)$ ≤$H(x)$ with equality iff $H(x|y)$=0 which is true only if $x$ is a function of $y$, i.e. if $y$ is a one-to-one function of $x$ for every value of $x$ with $p(x)$>0. Hence

   (a) $H(y)$≤$H(x)$ because, for example $1^2 = -1^2$

   (b) $H(y)$=$H(x)$

3. Maximum is log $n$ iff all elements of $p$ are equal. Minimum is 0 iff only one element of $p$ is non-zero; there are $n$ possible elements that this could be.

4. (a) and (b) are straightforward calculus: easiest to convert logs to base $e$ first. For the others, assume $½ < p < 1$ for covenience (other half follows by symmetry). Since $H''(p) < 0$, $H(p)$ is concave and so lies above the straight line $2-2p$ defined in (c).

   At $p = ½$ the bound in (e) has the same value and first two derivatives as $H(p)$. For $½ < p < 1$ its second derivative is greater than $H''(p)$ and so the bound follows.

   For (d) we consider $D(p) = H(p) - 1 + 4(p - ½)^2$. $D''(p) = 0$ is a quadratic in $p$ and has only two solutions $p = ½ \pm \sqrt{(2-\log e)/8} = 0.5 \pm 0.26$. Therefore $D'(p)$ increases from 0 at $p = 0.5$ to reach a maximum at $p = 0.76$ and decreases thereafter. This implies that $D'(p) = 0$ has only one solution for $p > ½$ and therefore that $D(p)$ has a single maximum. Since $D(½) = D(1) = 0$ we must have $D(p) > 0$ for $½ < p < 1$.

5. (a) chain rule, (b) $g(x)|x$ has only one possible value and hence zero entropy, (c) chain rule, (d) entropy is positive. We have equality at (d) iff $g(x)$ is a one-to-one function for every $x$ with $p(x)$>0.

6. $H(y \mid x) = \sum_x p(x) H(y \mid x = x)$

   All terms are non-negative so the sum is zero only if all terms are zero. For any given term this is true either if $p(x)$=0 or if H($y|x$=x) is zero. The second case arises only if H($y|x$=x) has only one value, i.e. $y$ is a function of $x$. The first case is why we needed the qualification about $p(x)$>0 in answers 2 and 4 above.

7. (a) The probability of any given value of $x_{1:4}$ depends on the number of 1's and 0's. We create four subsets with equal probabilities to generate a pair of bits and two other subsets to generate one bit only. The expected number of bits generated is

   $$E\,K = 8p(1-p)^3 + 10p^2(1-p)^2 + 8p^3(1-p)$$

   (b) (a) i.i.d entropies add, (b) functions reduce entropy, (c) chain rule, (d) $z_i$ are i.i.d. with entropy of 1 bit, (e) entropy is positive.

8. (a) This is true for any Markov chain $x \rightarrow y \rightarrow z$. One possibility is $x$=$y$=$z$ all fair Bernoulli variables.

   (b) An example of this was given in lectures. A slightly different example is if $x$ and $y$ are fair binary variables and $z$=$xy$. Knowing $z$, entangles $x$ and $y$.

9. (a) I($x;y;z$)={H($x$)-H($x|y$)}-{H($x|z$)-H($x|y,z$)}=H($x$)-{H($x,y$)-H($y$)}-{H($x,z$)-H($z$)}+{H($x,y,z$)-H($y,z$)}

   (b) Use the example from 7(b) above.

10. Define $f(x)$=ln($x$)+$x^{-1}$–1. This is continuous and differentiable in (0,∞). Differentiate twice to show that the only extremum occurs at $x$=1 and that it is a minimum. Hence f($x$)≥f(1)=0.

11. H($x$)=H($y$)=0.918; H($x|y$)=H($y|x$)=0.667; H($x,y$)=1.58; I($x;y$)=0.252.

12. The data processing inequality says that I($x;z$)≤I($x;y$)=H($y$)–H($y|x$) ≤H($y$) ≤ log k where the last inequality is the uniform bound on entropy. If $k$=1 then log $k$ = 0 and so $x$ and $z$ must be independent.

13. (a) $H(x,y|z)=H(x|z)+H(y|x,z)\geq H(x|z)$ with equality if $y$ is a function of $x$ and $z$.

(b) $I(x,y;z)=I(x;z)+I(y;z|x) \geq I(x;z)$ with equality if $y$ and $z$ are conditionally independent given $x$.

(c) $H(x,y,z)-H(x,y)=H(z|x,y)=H(z|x)-I(y;z|x) \leq H(z|x)=H(x,z)-H(x)$ with equality if $y$ and $z$ are conditionally independent given $x$.

(d) $I(x,y;z)=I(y;z)+I(x;z|y)= I(x;z)+I(y;z|x)$. Rearrange this to give the inequality which is in fact always an equality (trick question).