ECE 1502 — Information Theory

Problem Set 1 solutions[1]

October 5, 2007

2.1 (a) The number X of tosses till the first head appears has the geometric distribution with parameter $p = 1/2$, where $P(X = n) = pq^{n-1}$, $n \in \{1, 2, \ldots\}$. Hence the entropy of X is

$$\begin{aligned} H(X) &= -\sum_{n=1}^{\infty} pq^{n-1} \log(pq^{n-1}) \\ &= -\left[\sum_{n=0}^{\infty} pq^n \log p + \sum_{n=0}^{\infty} npq^n \log q\right] \\ &= \frac{-p \log p}{1-q} - \frac{pq \log q}{p^2} \\ &= \frac{-p \log p - q \log q}{p} \\ &= H(p)/p \text{ bits.} \end{aligned}$$

If $p = 1/2$, then $H(X) = 2$ bits.

(b) Intuitively, it seems clear that the best questions are those that have equally likely chances of receiving a yes or a no answer. Consequently, one possible guess is that the most "efficient" series of questions is: Is $X = 1$? If not, is $X = 2$? If not, is $X = 3$? ... with a resulting expected number of questions equal to $\sum_{n=1}^{\infty} n(1/2^n) = 2$. This should reinforce the intuition that $H(X)$ is a measure of the uncertainty of X. Indeed in this case, the entropy is exactly the same as the average number of questions needed to define X, and in general $E(\#\text{ of questions}) \geq H(X)$. This problem has an interpretation as a source coding problem. Let 0 =no, 1 =yes, X =Source, and Y =Encoded Source. Then the set of questions in the above procedure can be written as a collection of (X, Y) pairs: $(1, 1)$, $(2, 01)$, $(3, 001)$, etc. . In fact, this intuitively derived code is the optimal (Huffman) code minimizing the expected number of questions.

2.2 Let $y = g(x)$. Then

$$p(y) = \sum_{x:\, y=g(x)} p(x).$$

Consider any set of x's that map onto a single y. For this set

$$\sum_{x:\, y=g(x)} p(x) \log p(x) \leq \sum_{x:\, y=g(x)} p(x) \log p(y) = p(y) \log p(y),$$

since log is a monotone increasing function and $p(x) \leq \sum_{x:\, y=g(x)} p(x) = p(y)$. Extending this argument to the entire range of X (and Y), we obtain

$$\begin{aligned} H(X) &= -\sum_{x} p(x) \log p(x) \\ &= -\sum_{y} \sum_{x:\, y=g(x)} p(x) \log p(x) \end{aligned}$$

[1] Solutions to problems from the text are supplied courtesy of Joy A. Thomas.

$$\begin{aligned} &\geq -\sum_y p(y)\log p(y) \\ &= H(Y), \end{aligned}$$

with equality iff g is one-to-one with probability one.

(a) $Y = 2^X$ is one-to-one and hence the entropy, which is just a function of the probabilities (and not the values of a random variable) does not change, i.e., $H(X) = H(Y)$.

(b) $Y = \cos(X)$ is not necessarily one-to-one. Hence all that we can say is that $H(X) \geq H(Y)$, with equality if cosine is one-to-one on the range of X.

2.3 We wish to find *all* probability vectors $\mathbf{p} = (p_1, p_2, \ldots, p_n)$ which minimize

$$H(\mathbf{p}) = -\sum_i p_i \log p_i.$$

Now $-p_i \log p_i \geq 0$, with equality iff $p_i = 0$ or 1. Hence the only possible probability vectors which minimize $H(\mathbf{p})$ are those with $p_i = 1$ for some i and $p_j = 0, j \neq i$. There are n such vectors, i.e., $(1, 0, \ldots, 0)$, $(0, 1, 0, \ldots, 0)$, $\ldots$, $(0, \ldots, 0, 1)$, and the minimum value of $H(\mathbf{p})$ is 0.

2.4 *Entropy of functions of a random variable.*

(a) $H(X, g(X)) = H(X) + H(g(X)|X)$ by the chain rule for entropies.

(b) $H(g(X)|X) = 0$ since for any particular value of X, g(X) is fixed, and hence $H(g(X)|X) = \sum_x p(x) H(g(X)|X = x) = \sum_x 0 = 0$.

(c) $H(X, g(X)) = H(g(X)) + H(X|g(X))$ again by the chain rule.

(d) $H(X|g(X)) \geq 0$, with equality iff X is a function of $g(X)$, i.e., $g(.)$ is one-to-one. Hence $H(X, g(X)) \geq H(g(X))$.

Combining parts (b) and (d), we obtain $H(X) \geq H(g(X))$.

2.5 *Zero Conditional Entropy.* Assume that there exists an x, say x_0 and two different values of y, say y_1 and y_2 such that $p(x_0, y_1) > 0$ and $p(x_0, y_2) > 0$. Then $p(x_0) \geq p(x_0, y_1) + p(x_0, y_2) > 0$, and $p(y_1|x_0)$ and $p(y_2|x_0)$ are not equal to 0 or 1. Thus

$$\begin{aligned} H(Y|X) &= -\sum_x p(x) \sum_y p(y|x) \log p(y|x) && (1) \\ &\geq p(x_0)(-p(y_1|x_0)\log p(y_1|x_0) - p(y_2|x_0)\log p(y_2|x_0)) && (2) \\ > &> 0, && (3) \end{aligned}$$

since $-t \log t \geq 0$ for $0 \leq t \leq 1$, and is strictly positive for t not equal to 0 or 1. Therefore the conditional entropy $H(Y|X)$ is 0 if and only if Y is a function of X.

2.8 *Drawing with and without replacement.* Intuitively, it is clear that if the balls are drawn with replacement, the number of possible choices for the i-th ball is larger, and therefore the conditional entropy is larger. But computing the conditional distributions is slightly involved. It is easier to compute the unconditional entropy.

- With replacement. In this case the conditional distribution of each draw is the same for every draw. Thus

$$X_i = \begin{cases} \text{red} & \text{with prob.} \frac{r}{r+w+b} \\ \text{white} & \text{with prob.} \frac{w}{r+w+b} \\ \text{black} & \text{with prob.} \frac{b}{r+w+b} \end{cases} \tag{4}$$

and therefore

$$H(X_i|X_{i-1},\ldots,X_1) = H(X_i) \tag{5}$$

$$= \log(r+w+b) - \frac{r}{r+w+b}\log r - \frac{w}{r+w+b}\log w - \frac{b}{r+w+b}\log b. \tag{6}$$

- Without replacement. The unconditional probability of the i-th ball being red is still $r/(r+w+b)$, etc. Thus the unconditional entropy $H(X_i)$ is still the same as with replacement. The conditional entropy $H(X_i|X_{i-1},\ldots,X_1)$ is strictly less than the unconditional entropy. This can be seen by recalling that $H(X_i|X_{i-1},\ldots,X_1) = H(X_i)$ if and only if X_i is independent of $\{X_{i-1},\ldots,X_1\}$. Clearly, without replacement, X_i is not independent of the past, and therefore the entropy of drawing without replacement is lower.

2.10 *Entropy.* We can do this problem by writing down the definition of entropy and expanding the various terms. Instead, we will use the algebra of entropies for a simpler proof.

Since X_1 and X_2 have disjoint support sets, we can write

$$X = \begin{cases} X_1 & \text{with probability} & \alpha \\ X_2 & \text{with probability} & 1-\alpha \end{cases}$$

Define a function of X,

$$\theta = f(X) = \begin{cases} 1 & \text{when } X = X_1 \\ 2 & \text{when } X = X_2 \end{cases}$$

Then as in problem 1, we have

$$\begin{aligned} H(X) &= H(X, f(X)) = H(\theta) + H(X|\theta) \\ &= H(\theta) + p(\theta=1)H(X|\theta=1) + p(\theta=2)H(X|\theta=2) \\ &= H(\alpha) + \alpha H(X_1) + (1-\alpha)H(X_2) \end{aligned}$$

where $H(\alpha) = -\alpha\log\alpha - (1-\alpha)\log(1-\alpha)$.

Finally, we have

$$\begin{aligned} H(X) &\le H(\alpha) + \max(H(X_1), H(X_2)) \\ &\le 1 + \max(H(X_1), H(X_2)). \end{aligned}$$

Without loss of generality, we assume $H(X_1) \ge H(X_2)$, then

$$H(X) \le 1 + H(X_1),$$

and

$$\begin{aligned} 2^{H(X)} &\leq 2^{1+H(X_1)} \\ &= 2^{H(X_1)} + 2^{H(X_1)} \\ &\leq 2^{H(X_1)} + 2^{H(X_2)} \end{aligned}$$

as desired. Since $H(X)$ is the expected number of bits required to represent an element of X, $2^{H(X)}$ is the number of symbols in the effective alphabet. In the context of a disjoint mixture, the preceding implies that the effective alphabet size of the mixture is smaller than the sum of the effective alphabet sizes of the individual inputs.

2.11 *A measure of correlation.* X_1 and X_2 are identically distributed and

$$\rho = 1 - \frac{H(X_2|X_1)}{H(X_1)}$$

(a)

$$\begin{aligned} \rho &= \frac{H(X_1) - H(X_2|X_1)}{H(X_1)} \\ &= \frac{H(X_2) - H(X_2|X_1)}{H(X_1)} \text{ (since } H(X_1) = H(X_2)) \\ &= \frac{I(X_1; X_2)}{H(X_1)}. \end{aligned}$$

(b) Since $0 \leq H(X_2|X_1) \leq H(X_2) = H(X_1)$, we have

$$0 \leq \frac{H(X_2|X_1)}{H(X_1)} \leq 1$$

$$0 \leq \rho \leq 1.$$

(c) $\rho = 0$ iff $I(X_1; X_2) = 0$ iff X_1 and X_2 are independent.

(d) $\rho = 1$ iff $H(X_2|X_1) = 0$ iff X_2 is a function of X_1. By symmetry, X_1 is a function of X_2, i.e., X_1 and X_2 have a one-to-one relationship.