ECE 542: Information Theory and Coding
Homework 1 Solutions
Problems 2.1, 2.2, 2.6, 2.8, 2.14, 2.21, 2.22, 2.30

1. *Coin flips.* 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. The following expressions may be useful:

$$\sum_{n=1}^{\infty} r^n = \frac{r}{1-r}, \qquad \sum_{n=1}^{\infty} nr^n = \frac{r}{(1-r)^2}.$$

   (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.

Solution:

   (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. *Entropy of functions.* Let X be a random variable taking on a finite number of values. What is the (general) inequality relationship of $H(X)$ and $H(Y)$ if

(a) $Y = 2^X$?

(b) $Y = \cos X$?

Solution: 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

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

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.

6. *Zero conditional entropy.* Show that if $H(Y|X) = 0$, then Y is a function of X, i.e., for all x with $p(x) > 0$, there is only one possible value of y with $p(x,y) > 0$.

Solution: *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

$$\begin{aligned} H(Y|X) &= -\sum_x p(x) \sum_y p(y|x) \log p(y|x) && (2.66) \\ &\geq p(x_0)(-p(y_1|x_0) \log p(y_1|x_0) - p(y_2|x_0) \log p(y_2|x_0)) && (2.67) \\ > &> 0, && (2.68) \end{aligned}$$

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.

8. *World Series.* The World Series is a seven-game series that terminates as soon as either team wins four games. Let X be the random variable that represents the outcome of a World Series between teams A and B; possible values of X are AAAA, BABABAB, and BBBAAAA. Let Y be the number of games played, which ranges from 4 to 7. Assuming that A and B are equally matched and that the games are independent, calculate $H(X)$, $H(Y)$, $H(Y|X)$, and $H(X|Y)$.

Solution:

World Series. Two teams play until one of them has won 4 games.

There are 2 (AAAA, BBBB) World Series with 4 games. Each happens with probability $(1/2)^4$.

There are $8 = 2\binom{4}{3}$ World Series with 5 games. Each happens with probability $(1/2)^5$.

There are $20 = 2\binom{5}{3}$ World Series with 6 games. Each happens with probability $(1/2)^6$.

There are $40 = 2\binom{6}{3}$ World Series with 7 games. Each happens with probability $(1/2)^7$.

The probability of a 4 game series ($Y = 4$) is $2(1/2)^4 = 1/8$.

The probability of a 5 game series ($Y = 5$) is $8(1/2)^5 = 1/4$.

The probability of a 6 game series ($Y = 6$) is $20(1/2)^6 = 5/16$.

The probability of a 7 game series ($Y = 7$) is $40(1/2)^7 = 5/16$.

$$\begin{aligned} H(X) &= \sum p(x) log \frac{1}{p(x)} \\ &= 2(1/16) \log 16 + 8(1/32) \log 32 + 20(1/64) \log 64 + 40(1/128) \log 128 \\ &= 5.8125 \end{aligned}$$

$$
\begin{aligned}
H(Y) &= \sum p(y) log \frac{1}{p(y)} \\
&= 1/8 \log 8 + 1/4 \log 4 + 5/16 \log(16/5) + 5/16 \log(16/5) \\
&= 1.924
\end{aligned}
$$

Y is a deterministic function of X, so if you know X there is no randomness in Y. Or, $H(Y|X) = 0$.

Since $H(X) + H(Y|X) = H(X,Y) = H(Y) + H(X|Y)$, it is easy to determine $H(X|Y) = H(X) + H(Y|X) - H(Y) = 3.889$

14. *Drawing with and without replacement.* An urn contains r red, w white, and b black balls. Which has higher entropy, drawing $k \geq 2$ balls from the urn with replacement or without replacement? Set it up and show why. (There is both a hard way and a relatively simple way to do this.)

    Solution: *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

    $$
    X_i = \begin{cases} \text{red} & \text{with prob.} \frac{r}{r+w+b} \\ \text{white} & \text{with prob.} \frac{w}{r+w+b} \\ \text{black} & \text{with prob.} \frac{b}{r+w+b} \end{cases} \tag{2.83}
    $$

    and therefore

    $$
    H(X_i|X_{i-1}, \ldots, X_1) = H(X_i) \tag{2.84}
    $$

    $$
    = \log(r+w+b) - \frac{r}{r+w+b}\log r - \frac{w}{r+w+b}\log w - \frac{b}{r+w+b}\log b \tag{2.85}
    $$

    - 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 less than the unconditional entropy, and therefore the entropy of drawing without replacement is lower.

21. *Data processing.* Let $X_1 \to X_2 \to X_3 \to \cdots \to X_n$ form a Markov chain in this order; i.e., let

$$p(x_1, x_2, \ldots, x_n) = p(x_1)p(x_2|x_1)\cdots p(x_n|x_{n-1}).$$

Reduce $I(X_1; X_2, \ldots, X_n)$ to its simplest form.

Solution: *Data Processing.* By the chain rule for mutual information,

$$I(X_1; X_2, \ldots, X_n) = I(X_1; X_2) + I(X_1; X_3|X_2) + \cdots + I(X_1; X_n|X_2, \ldots, X_{n-2}). \quad (2.95)$$

By the Markov property, the past and the future are conditionally independent given the present and hence all terms except the first are zero. Therefore

$$I(X_1; X_2, \ldots, X_n) = I(X_1; X_2). \quad (2.96)$$

22. *Bottleneck.* Suppose a (non-stationary) Markov chain starts in one of n states, necks down to $k < n$ states, and then fans back to $m > k$ states. Thus $X_1 \to X_2 \to X_3$, $X_1 \in \{1, 2, \ldots, n\}$, $X_2 \in \{1, 2, \ldots, k\}$, $X_3 \in \{1, 2, \ldots, m\}$.

   (a) Show that the dependence of X_1 and X_3 is limited by the bottleneck by proving that $I(X_1; X_3) \leq \log k$.

   (b) Evaluate $I(X_1; X_3)$ for $k = 1$, and conclude that no dependence can survive such a bottleneck.

Solution:

Bottleneck.

(a) From the data processing inequality, and the fact that entropy is maximum for a uniform distribution, we get

$$\begin{aligned} I(X_1; X_3) &\leq I(X_1; X_2) \\ &= H(X_2) - H(X_2 \mid X_1) \\ &\leq H(X_2) \\ &\leq \log k. \end{aligned}$$

Thus, the dependence between X_1 and X_3 is limited by the size of the bottleneck. That is $I(X_1; X_3) \leq \log k$.

(b) For $k = 1$, $I(X_1; X_3) \leq \log 1 = 0$ and since $I(X_1, X_3) \geq 0$, $I(X_1, X_3) = 0$. Thus, for $k = 1$, X_1 and X_3 are independent.

30)

Recall that,

$$-\sum_{i=0}^{\infty} p_i \log p_i \leq -\sum_{i=0}^{\infty} p_i \log q_i.$$

Let $q_i = \alpha(\beta)^i$. Then we have that,

$$\begin{aligned}
-\sum_{i=0}^{\infty} p_i \log p_i &\leq -\sum_{i=0}^{\infty} p_i \log q_i \\
&= -\left(\log(\alpha)\sum_{i=0}^{\infty} p_i + \log(\beta)\sum_{i=0}^{\infty} i p_i\right) \\
&= -\log\alpha - A\log\beta
\end{aligned}$$

Notice that the final right hand side expression is independent of $\{p_i\}$, and that the inequality,

$$-\sum_{i=0}^{\infty} p_i \log p_i \leq -\log\alpha - A\log\beta$$

holds for all α, β such that,

$$\sum_{i=0}^{\infty} \alpha\beta^i = 1 = \alpha\frac{1}{1-\beta}.$$

The constraint on the expected value also requires that,

$$\sum_{i=0}^{\infty} i\alpha\beta^i = A = \alpha\frac{\beta}{(1-\beta)^2}.$$

Combining the two constraints we have,

$$\begin{aligned}
\alpha\frac{\beta}{(1-\beta)^2} &= \left(\frac{\alpha}{1-\beta}\right)\left(\frac{\beta}{1-\beta}\right) \\
&= \frac{\beta}{1-\beta} \\
&= A,
\end{aligned}$$

which implies that,

$$\begin{aligned}
\beta &= \frac{A}{A+1} \\
\alpha &= \frac{1}{A+1}.
\end{aligned}$$

So the entropy maximizing distribution is,

$$p_i = \frac{1}{A+1}\left(\frac{A}{A+1}\right)^i.$$

Plugging these values into the expression for the maximum entropy,

$$-\log \alpha - A \log \beta = (A+1)\log(A+1) - A \log A.$$

The general form of the distribution,

$$p_i = \alpha \beta^i$$

can be obtained either by guessing or by Lagrange multipliers where,

$$F(p_i, \lambda_1, \lambda_2) = -\sum_{i=0}^{\infty} p_i \log p_i + \lambda_1 (\sum_{i=0}^{\infty} p_i - 1) + \lambda_2 (\sum_{i=0}^{\infty} i p_i - A)$$

is the function whose gradient we set to 0.

Many of you used Lagrange multipliers, but failed to argue that the result obtained is a global maximum. An argument similar to the above should have been used. On the other hand one could simply argue that since $-H(p)$ is convex, it has only one local minima, no local maxima and therefore Lagrange multiplier actually gives the global maximum for $H(p)$.