1. (8 pts)

(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

$$H(X) = - \sum_{n=1}^{\infty} pq^{n-1} \log(pq^{n-1})$$

$$= - \left[ \sum_{n=0}^{\infty} pq^n \log p + \sum_{n=0}^{\infty} npq^n \log q \right]$$

$$= \frac{-p \log p}{1-q} - \frac{pq \log q}{p^2}$$

$$= \frac{-p \log p - q \log q}{p}$$

$$= H(p)/p \text{ bits.}$$

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. (8 pts) Mutual Information and the Weather

(a) \[ I(P_s; W) = H(P_s) - H(P_s|W) = 0 - 0 = 0 \] (1)

(b) \[ I(P_w; W) = H(W) - H(W|P_w) = H(.9) - .75 \times H(0) - .25 \times H(.4) = 0.2262 \] (2)

(c) Wendy, provides the most information. In fact, Stormy provides no information at all.

(d) Plant your tulip bulbs when Wendy forecasts rain.

3. (4 pts) 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.

4. (8 pts) 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

\[ X_i = \begin{cases} \text{red} & \text{with prob. } \frac{r}{r+w+b} \\ \text{white} & \text{with prob. } \frac{w}{r+w+b} \\ \text{black} & \text{with prob. } \frac{b}{r+w+b} \end{cases} \] (3)

and therefore

\[ H(X_i|X_{i-1}, \ldots, X_1) = H(X_i) \] (4)

\[ = \log(r + w + b) - \frac{r}{r+w+b} \log r - \frac{w}{r+w+b} \log w - \frac{b}{r+w+b} \log b. \] (5)

- 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.
5. (12 pts) *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 \( H(X) \geq H(g(X)) \).

6. (12 pts) *Example of joint entropy*

(a) \( H(X) = \frac{2}{3} \log \frac{3}{2} + \frac{1}{3} \log 3 = 0.918 \) bits = \( H(Y) \).

(b) \( H(X|Y) = \frac{1}{3}H(X|Y = 0) + \frac{2}{3}H(X|Y = 1) = 0.667 \) bits = \( H(Y|X) \).

(c) \( H(X, Y) = 3 \times \frac{1}{3} \log 3 = 1.585 \) bits.

(d) \( H(Y) - H(Y|X) = 0.251 \) bits.

(e) \( I(X;Y) = H(Y) - H(Y|X) = 0.251 \) bits.

(f) See Figure 2.2 in *Elements of Information Theory.*

7. (15 pts)

(a) For convenience in notation, we will let

\[
S_k = \sum_{i=1}^{k} p_i
\]

and we will denote \( H_2(q, 1 - q) \) as \( h(q) \). Then we can write the grouping axiom as

\[
H_m(p_1, \ldots, p_m) = H_{m-1}(S_2, p_3, \ldots, p_m) + S_2 h \left( \frac{p_2}{S_2} \right).
\]

(7)

Applying the grouping axiom again, we have

\[
H_m(p_1, \ldots, p_m) = H_{m-1}(S_2, p_3, \ldots, p_m) + S_2 h \left( \frac{p_2}{S_2} \right) = H_{m-2}(S_3, p_4, \ldots, p_m) + S_3 h \left( \frac{p_3}{S_3} \right) + S_2 h \left( \frac{p_2}{S_2} \right)
\]

(8)

\[
\vdots
\]

(10)

\[
= H_{m-(k-1)}(S_k, p_{k+1}, \ldots, p_m) + \sum_{i=2}^{k} S_i h \left( \frac{p_i}{S_i} \right).
\]

(11)
Now, we apply the same grouping axiom repeatedly to $H_k(p_1/S_k, \ldots, p_k/S_k)$, to obtain

$$H_k\left(\frac{p_1}{S_k}, \ldots, \frac{p_k}{S_k}\right) = H_2\left(\frac{S_{k-1}}{S_k}, \frac{p_k}{S_k}\right) + \sum_{i=2}^{k-1} \frac{S_i}{S_k} h\left(\frac{p_i}{S_i/S_k}\right)$$

(12)

$$= \frac{1}{S_k} \sum_{i=2}^{k} S_i h\left(\frac{p_i}{S_i}\right).$$

(13)

From (11) and (13), it follows that

$$H_m(p_1, \ldots, p_m) = H_{m-k+1}(S_k, p_{k+1}, \ldots, p_m) + S_k H_{k}\left(\frac{p_1}{S_k}, \ldots, \frac{p_k}{S_k}\right),$$

(14)

which is the extended grouping axiom.

(b) We need to use an axiom that is not explicitly stated in the text, namely that the function $H_m$ is symmetric with respect to its arguments. Using this, we can combine any set of arguments of $H_m$ using the extended grouping axiom.

Consider

$$f(mn) = H_{mn}\left(\frac{1}{mn}, \frac{1}{mn}, \ldots, \frac{1}{mn}\right).$$

(15)

By repeatedly applying the extended grouping axiom, we have

$$f(mn) = H_{mn}\left(\frac{1}{mn}, \frac{1}{mn}, \ldots, \frac{1}{mn}\right)$$

(16)

$$= H_{mn-n}\left(\frac{1}{m}, \frac{1}{mn}, \ldots, \frac{1}{mn}\right) + \frac{1}{m} H_n\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$$

(17)

$$= H_{mn-2n}\left(\frac{1}{m}, \frac{1}{m}, \frac{1}{mn}, \ldots, \frac{1}{mn}\right) + \frac{2}{m} H_n\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$$

(18)

$$\vdots$$

(19)

$$= H_m\left(\frac{1}{m}, \ldots, \frac{1}{m}\right) + H\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$$

(20)

$$= f(m) + f(n).$$

(21)

We can immediately use this to conclude that $f(m^k) = kf(m)$.

(c) Note that in this problem only, the subscript of $H$ is not the base of the logarithm but rather the number of arguments of $H$. Another way of stating the original Normalization Axiom is $H_2(\frac{1}{2}, \frac{1}{2}) = \log_2(2)$. The new Normalization Axiom is $H_2(\frac{1}{2}, \frac{1}{2}) = \log_e(2) = \ln(2)$. 

4
8. (10 pts) Conditional mutual information vs. unconditional mutual information.

(a) The last corollary to Theorem 2.8.1 in the text states that if $X \rightarrow Y \rightarrow Z$ that is, if $p(x, y \mid z) = p(x \mid z)p(y \mid z)$ then, $I(X;Y) \geq I(X;Y \mid Z)$. Equality holds if and only if $I(X;Z) = 0$ or $X$ and $Z$ are independent.

A simple example of random variables satisfying the inequality conditions above is, $X$ is a fair binary random variable and $Y = X$ and $Z = Y$. In this case, $I(X;Y) = H(X) - H(X \mid Y) = H(X) = 1$ and, $I(X;Y \mid Z) = H(X \mid Z) - H(X \mid Y,Z) = 0$. So that $I(X;Y) > I(X;Y \mid Z)$.

(b) This example is also given in the text. Let $X, Y$ be independent fair binary random variables and let $Z = X + Y$. In this case we have that, $I(X;Y) = 0$ and, $I(X;Y \mid Z) = H(X \mid Z) = 1/2$. So $I(X;Y) < I(X;Y \mid Z)$. Note that in this case $X, Y, Z$ are not markov.
9. (10 pts) *Coin weighing.*

(a) For $n$ coins, there are $2n + 1$ possible situations or “states”.

- One of the $n$ coins is heavier.
- One of the $n$ coins is lighter.
- They are all of equal weight.

Each weighing has three possible outcomes - equal, left pan heavier or right pan heavier. Hence with $k$ weighings, there are $3^k$ possible outcomes and hence we can distinguish between at most $3^k$ different “states”. Hence $2n + 1 \leq 3^k$ or $n \leq (3^k - 1)/2$.

Looking at it from an information theoretic viewpoint, each weighing gives at most $\log_2 3$ bits of information. There are $2n + 1$ possible “states”, with a maximum entropy of $\log_2(2n + 1)$ bits. Hence in this situation, one would require at least $\frac{\log_2(2n + 1)}{\log_2 3}$ weighings to extract enough information for determination of the odd coin, which gives the same result as above.

(b) The intuition is to try to make each of the three outcomes of a weighing approximately equally likely. A correct strategy is as follows:

- Set aside coins 9-12 and weigh the coins 1-8, with 1-4 on the left and 5-8 on the right.
- If the scale balances, we focus our attention on 9-12, otherwise we focus our attention on 1-8.
- Suppose the first weighing balances. Set aside 1-4, 6-8, and also 12. Place (known good) 5 and 11 on the left and on the right place 9-10. Now, if the scale balances, we can easily check 12 in a final weighing with 12 on the right and (say) 1 on the left. If the scale tips we should do a final weighing with 9 on the left and 10 on the right.
- Suppose the first weighing tips. Set aside 1-3, place 4-6 on the left and 7-9 on the right. Whether the scale balances casting suspicion on 1-3 or tips either way casting suspicion on either 4,7-8 or 5-6, there will be at most three potentially bent coins. For each we already know that it either can’t be heavy or can’t be light. A final weighing of only two of the three possibly suspect coins will suffice.

One question from the office hour was whether the bound derived in part (a) was actually achievable. For example, can one distinguish 13 coins in 3 weighings? No, not with a scheme like the one above. Yes, under the assumptions under which the bound was derived. The bound did not prohibit the division of coins into halves, neither did it disallow the existence of another coin known to be normal. Under both these conditions, it is possible to find the odd coin of 13 coins in 3 weighings. You could try modifying the above scheme to these cases.