Solutions to HW Set # 1

1. *Coin flips.*

   (a) The number X of tosses till the first head appears has a 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 $X = 3$? And so on, 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.* Suppose $X \sim P$ on A, and let $y = g(x)$. Then the probability mass function of Y satisfies

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

   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 $P(x) \le \sum_{x:\, y=g(x)} P(x) = P(y)$. Extending this argument to the entire range of X (and Y), we obtain

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

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

In the first case, $Y = 2^X$ 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., $H(X) = H(Y)$.

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

For part (ii), we have $H(X, g(X)) = H(X) + H(g(X)|X)$ by the chain rule for entropy. Then $H(g(X)|X) = 0$, since, for any particular value of X, g(X) is fixed, and hence $H(g(X)|X) = \sum_x p(x)H(g(X)|X = x) = \sum_x 0 = 0$. Similarly, $H(X, g(X)) = H(g(X)) + H(X|g(X))$ again by the chain rule. And finally, $H(X|g(X)) \ge 0$, with equality iff X is a function of $g(X)$, i.e., g is one-to-one (why?). Hence $H(X, g(X)) \ge H(g(X))$.

3. *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 $P(x_0) \ge P(x_0, y_1) + P(x_0, y_2) > 0$, and $P(y_1|x_0)$ and $P(y_2|x_0)$ are not equal to 0 or 1. Thus

$$\begin{aligned} H(Y|X) &= -\sum_x P(x) \sum_y P(y|x)\log P(y|x) \\ &\ge P(x_0)(-P(y_1|x_0)\log P(y_1|x_0) - P(y_2|x_0)\log P(y_2|x_0)) \\ &> 0, \end{aligned}$$

since $-t\log t \ge 0$ for $0 \le t \le 1$, and is strictly positive for t not equal to 0 or 1. Therefore, the conditional entropy $H(Y|X)$ is 0 only if Y is a function of X. The converse (the "if" part) is trivial (why?).

4. *Entropy of a disjoint mixture.* 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

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

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

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

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

The maximization over α and the resulting inequality is simple calculus. The interesting point here is the following: From the AEP we know that, instead of considering all $|A|^n$ strings, we can concentrate on the $\approx 2^{nH} = (2^H)^n$ typical strings. In other words, we can pretend we have a "completely random," or uniform source, with alphabet size $2^H < |A|$, so the effective alphabet size of X is not $|A$, but $2^{H(X)}$.

The inequality we get here says that the effective alphabet size of the mixture X of the random variables X_1, X_2 is no larger than the sum of their effective alphabet sizes.

5. *Run length coding.* Since the run lengths are a function of X_1^n, $H(R) \le H(X_1^n)$. Any X_i together with the run lengths determine the entire sequence X_1^n. Hence

$$\begin{aligned} H(X_1^n) &= H(X_i, R) \\ &= H(R) + H(X_i|R) \\ &\le H(R) + H(X_i) \\ &\le H(R) + 1. \end{aligned}$$

6. *Markov's inequality for probabilities.* We have:

$$\begin{aligned} \Pr(P(X) < d) \log \frac{1}{d} &= \sum_{x:P(x)<d} P(x) \log \frac{1}{d} \\ &\le \sum_{x:P(x)<d} P(x) \log \frac{1}{P(x)} \\ &\le \sum_{x} P(x) \log \frac{1}{P(x)} \\ &= H(X). \end{aligned}$$

7. *The AEP and source coding.*

(a) The number of 100-bit binary sequences with three or fewer ones is:

$$\binom{100}{0} + \binom{100}{1} + \binom{100}{2} + \binom{100}{3} = 1 + 100 + 4950 + 161700 = 166751\,.$$

The required codeword length is $\lceil \log_2 166751 \rceil = 18$. (Note that $h(0.005) \approx 0.0454$, so 18 is quite a bit larger than the optimal $100 \times h(0.005) \approx 4.5$ bits of entropy.)

(b) The probability that a 100-bit sequence has three or fewer ones is:

$$\sum_{i=0}^{3} \binom{100}{i} (0.005)^i (0.995)^{100-i} \approx 0.60577 + 0.30441 + 0.7572 + 0.01243 = 0.99833.$$

Thus, the probability that the sequence that is generated cannot be encoded is $\approx 1 - 0.99833 = 0.00167$.

(c) If S_n that is the sum of$n$ IID random variables $X_1, X_2, \ldots, X_n$, Chebyshev's inequality states that,

$$\Pr(|S_n - n\mu| \geq \epsilon) \leq \frac{n\sigma^2}{\epsilon^2},$$

where μ and σ^2 are the mean and variance of the X_i. (Therefore $n\mu$ and $n\sigma^2$ are the mean and variance of S_n.) In this problem, $n = 100$, $\mu = 0.005$, and $\sigma^2 = (0.005)(0.995)$. Note that $S_{100} \geq 4$ if and only if $|S_{100} - 100(0.005)| \geq 3.5$, so we should choose $\epsilon = 3.5$. Then,

$$\Pr(S_{100} \geq 4) \leq \frac{100(0.005)(0.995)}{(3.5)^2} \approx 0.04061\,.$$

This bound is much larger than the actual probability 0.00167.

8. Since the X_1^n are IID, so are $Q(X_1), Q(X_2), \ldots, Q(X_n)$, and hence we can apply the (weak or strong, depending on your preference) law of large numbers to obtain,

$$\begin{aligned}
\lim -\frac{1}{n} \log Q^n(X_1^n) &= \lim -\frac{1}{n} \sum \log Q(X_i) \\
&= E[-\log Q(X_1)] \qquad \text{[in probability, or w.p. 1]} \\
&= -\sum_x P(x) \log Q(x) \\
&= \sum_x P(x) \log \frac{P(x)}{Q(x)} - \sum_x P(x) \log P(x) \\
&= D(P\|Q) + H(P).
\end{aligned}$$

9. *Random box size.* The volume $V_n = \prod_{i=1}^n X_i$ is a random variable. Since the X_i are random variables uniformly distributed on $[0,1]$, we expect that V_n tends to 0 as $n \to \infty$. However,

$$\log_e V_n^{\frac{1}{n}} = \frac{1}{n} \log_e V_n = \frac{1}{n} \sum \log_e X_i \to E(\log_e(X)) \quad \text{in probability,}$$

by the weak law of large numbers, since the RVs $\log_e(X_i)$ are IID. Now,

$$E[\log_e(X_i)] = \int_0^1 \log_e(x)\,dx = -1.$$

Hence, since e^x is a continuous function,

$$\lim_{n\to\infty} V_n^{\frac{1}{n}} = e^{\lim_{n\to\infty} \frac{1}{n}\log_e V_n} = \frac{1}{e} < \frac{1}{2}.$$

Thus the "effective" edge length of this solid is e^{-1}. Note that since the X_i's are independent, $E(V_n) = \prod E(X_i) = (\frac{1}{2})^n$. [Also $\frac{1}{2}$ is the arithmetic mean of the random variables, and $\frac{1}{e}$ is their geometric mean.]