2nd Semester 2010

Solutions to Set #3 Part B
Data Compression via Matlab, AEP and block source coding

1. **Matlab simulation of Compression**

   Recall Question 3 from part A of HW 3

   Give a Huffman encoding into an alphabet of size D = 2 of the following probability mass function:

$$P_X = \left(\frac{1}{2}, \frac{1}{8}, \frac{1}{8}, \frac{1}{16}, \frac{1}{16}, \frac{1}{16}, \frac{1}{16}\right)$$

   Assume you have a file of size 10,000 symbols where the symbols are distributed i.i.d. according to the pmf above. After applying the Huffman code, what would be the pmf of the compressed binary file (namely, what is the probability of '0' and '1' in the compressed file), and what would be the expected length?

   (a) Generate a sequence (using Matlab or any other software) of 10,000 symbols of X with i.i.d. probability P_X. Assume the alphabet of X is $\mathcal{X} = (0, 1, ..., 6)$.

   (b) What is the percentage of each symbol $(0, 1, ..., 6)$ in the sequence. Explain the result this using the law of large numbers.

   (c) Represent each symbol in $\mathcal{X}$ using a simple binary representation. Namely, $X = 0$ represent as '000', $X = 1$ represent as '001',$X = 2$ represent as '010',..., $X = 6$ represent as '110'.

   (d) What is the length of the simple representation. What percentage of '0' and '1' do you have in this representation?

   (e) Now, compress the 10,000 symbols of X, into bits using Huffman code.

   (f) What is the length of the compressed file. What percentage of '0' and '1' do you have in this representation?

   (g) Explain the results using the law of large number and the analytical solution of Question 3 from HW 3.

2. **An AEP-like limit.** Let $X_1, X_2, \ldots$ be i.i.d. drawn according to probability mass function $p(x)$. Find

$$\lim_{n\to\infty} [p(X_1, X_2, \ldots, X_n)]^{\frac{1}{n}}.$$

Solution: An AEP-like limit.
$X_1, X_2, \ldots$, i.i.d. $\sim p(x)$. Hence $\log(X_i)$ are also i.i.d. and

$$\begin{aligned}
\lim(p(X_1, X_2, \ldots, X_n))^{\frac{1}{n}} &= \lim 2^{\log(p(X_1,X_2,\ldots,X_n))^{\frac{1}{n}}} \\
&= 2^{\lim \frac{1}{n}\sum \log p(X_i)} \\
&= 2^{E(\log(p(X)))} \\
&= 2^{-H(X)}
\end{aligned}$$

by the strong law of large numbers.

3. **AEP**. Let $X_1, X_2, \ldots$ be independent identically distributed random variables drawn according to the probability mass function $p(x), x \in \{1, 2, \ldots, m\}$. Thus $p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p(x_i)$. We know that $-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) \to H(X)$ in probability. Let $q(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} q(x_i)$, where q is another probability mass function on $\{1, 2, \ldots, m\}$.

   (a) Evaluate $\lim -\frac{1}{n}\log q(X_1, X_2, \ldots, X_n)$, where $X_1, X_2, \ldots$ are i.i.d. $\sim p(x)$.

   (b) Now evaluate the limit of the log likelihood ratio $\frac{1}{n}\log\frac{q(X_1,\ldots,X_n)}{p(X_1,\ldots,X_n)}$ when $X_1, X_2, \ldots$ are i.i.d. $\sim p(x)$. Thus the odds favouring q are exponentially small when p is true.

Solution: AEP.

   (a) Since the $X_1, X_2, \ldots, X_n$ are i.i.d., so are $q(X_1), q(X_2), \ldots, q(X_n)$,

and hence we can apply the strong law of large numbers to obtain

$$\begin{aligned}
\lim -\frac{1}{n}\log q(X_1, X_2, \ldots, X_n) &= \lim -\frac{1}{n}\sum \log q(X_i) \\
&= -E(\log q(X)) \text{ w.p. } 1 \\
&= -\sum p(x)\log q(x) \\
&= \sum p(x)\log\frac{p(x)}{q(x)} - \sum p(x)\log p(x) \\
&= D(\mathbf{p}||\mathbf{q}) + H(\mathbf{p}).
\end{aligned}$$

(b) Again, by the strong law of large numbers,

$$\begin{aligned}
\lim -\frac{1}{n}\log\frac{q(X_1, X_2, \ldots, X_n)}{p(X_1, X_2, \ldots, X_n)} &= \lim -\frac{1}{n}\sum \log\frac{q(X_i)}{p(X_i)} \\
&= -E(\log\frac{q(X)}{p(X)}) \text{ w.p. } 1 \\
&= -\sum p(x)\log\frac{q(x)}{p(x)} \\
&= \sum p(x)\log\frac{p(x)}{q(x)} \\
&= D(\mathbf{p}||\mathbf{q}).
\end{aligned}$$

4. **Lossless source coding with side information.**
Consider the lossless source coding with side information that is available at the encoder and decoder, where the source X and the side information Y are i.i.d. $\sim P_{X,Y}(x, y)$.

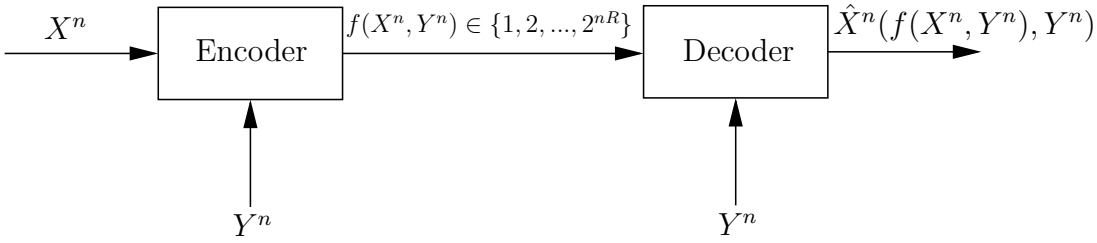

Figure 1: Lossless source coding with side information at the encoder and decoder.

Show that a code with rate $R < H(X|Y)$ can not be achievable, and interpret the result.

Hint: Let $T \triangleq f(X^n, Y^n)$. Consider

$$\begin{aligned} nR &\geq H(T) \\ &\geq H(T|Y^n), \end{aligned} \tag{1}$$

and use similar steps, including Fano's inequality, as we used in the class to prove the converse where side information was not available.

Solution Sketch of the solution (please fill in the explanation for each step):

$$\begin{aligned} nR &\geq H(T) \\ &\geq H(T|Y^n), \\ &\geq I(X^n; T|Y^n) \\ &= H(X^n|Y^n) - H(X^n|T, Y^n) \\ &= nH(X|Y) - \epsilon_n, \end{aligned}$$

where $\epsilon_n \to 0$.