ENEE 627
SPRING 2011
INFORMATION THEORY

ANSWER KEY TO TEST # 1:

1. _____

1.a. As shown in class, $H_2(\boldsymbol{u}) = H_2(M) = \log_2 M$.

1.b. The optimal code $C^\star_{2^L} : \{1, \ldots, 2^L\} \to \{0,1\}^\star$ is the one that corresponds to the full tree with 2^L terminal nodes (labelled $1, \ldots, 2^L$). Every codeword having length L, this code is indeed optimal since its average code length coincides with the entropy of the source, namely $H_2(M) = \log_2 M = \log_2 \left(2^L\right) = L$. One way to describe $C^\star_{2^L}$ is as follows: For each $m = 1, \ldots, 2^L$, write $C^\star_{2^L}(m)$ as the L-bit binary expansion of $m-1$.

1.c. Consider now the general case $M = 2^L + K$ with integers L and K satisfying

$$L = 1, 2, \ldots \quad \text{and} \quad K = 0, \ldots, 2^L - 1.$$

With code $C^\star_{2^L} : \{1, \ldots, 2^L\} \to \{0,1\}^\star$ described in Part **1.b**, we first set

$$C^\star_M(m) = C^\star_{2^L}(m), \quad m = 1, \ldots 2^L - K.$$

Next, on the remaining range $m = 2^L - K + 1, \ldots, 2^L + K$, group the symbols in pairs, say

$$2^L - K + 2k + 1 \quad \text{and} \quad 2^L - K + 2(k+1), \quad k = 0, 1, \ldots, K-1.$$

We now define the corresponding codewords by

$$C^\star_M(2^L - K + 2k + 1) = [C^\star_{2^L}(2^L - K + k), 0]$$

and

$$C^\star_M(2^L - K + 2(k+1)) = [C^\star_{2^L}(2^L - K + k), 1].$$

This corresponds to the following procedure (up to a relabeling of the nodes): Starting with the full binary tree associated with $C^\star_{2^L}$ (for the alphabet $\{1, \ldots, 2^L\}$), keep the terminal nodes corresponding to the symbols $m = 1, \ldots, 2^L - K$, and at the remaining nodes (which correspond to the symbols $m = 2^L - K + 1, \ldots, 2^L$) for the alphabet $\{1, \ldots, 2^L\}$), extend the tree by adding its two siblings.

It is immediate that $M-K$ codewords have length L, and that $2K$ codewords have length $L+1$, so that

$$\begin{aligned} L(M) &= \frac{1}{M}\left((2^L-K)L+2K(L+1)\right) \\ &= \frac{1}{M}\left((2^L+K)L+2K\right) \\ &= L+2\left(\frac{K}{M}\right). \end{aligned} \tag{1.1}$$

1.d. To have $L(M)=H_2(M)$ means that the optimal code $C_M^\star$ achieves the entropy bound. However, we know that this happens if and only if the underlying pmf is D-adic (here with $D=2$), namely

$$\frac{1}{M}=2^{-m(x)}, \quad x=1,\ldots,M$$

with positive integers $m(1),\ldots,m(M)$. Obviously this requires $m(1)=\ldots=m(M)=m^\star$ with $m^\star$ determined by

$$\frac{1}{M}=2^{-m^\star}.$$

Put another way, the equality $L(M)=H_2(M)$ requires that M be a power of two!

2. ---

2.a.

String x	$C(x)$	Probability $p(x)$
1	1000	λ
01	1001	$(1-\lambda)\lambda$
001	1010	$(1-\lambda)^2\lambda$
0001	1011	$(1-\lambda)^3\lambda$
00001	1100	$(1-\lambda)^4\lambda$
000001	1101	$(1-\lambda)^5\lambda$
0000001	1110	$(1-\lambda)^6\lambda$
00000001	1111	$(1-\lambda)^7\lambda$
00000000	0	$(1-\lambda)^8$

The procedure outlined in the problem statement can be intepreted as a binary code C for an i.i.d. source $(X,\boldsymbol{p})$ drawn from some alphabet $\mathcal{X}$ with

$$\mathcal{X}=\{1,01,001,0001,00001,000001,0000001,00000001,00000000\}.$$

This is so because the binary digits $\{Y_n,\ n=1,2,\ldots\}$ being i.i.d. rvs, the output $\{X_k,\ k=1,2,\ldots\}$ produced by the new source is also a sequence of i.i.d. rvs.

2.b. The code $C : \mathcal{X} \to \{0,1\}^\star$ is uniquely decodable because it is clearly a prefix code.

2.c. It is plain that

$$\begin{aligned}\mathbb{E}\left[\ell_C(X)\right] &= 1\mathbb{P}\left[X = 00000000\right] + 4\left(1 - \mathbb{P}\left[X = 00000000\right]\right) \\ &= (1-\lambda)^8 + 4\left(1 - (1-\lambda)^8\right) \\ &= 4 - 3(1-\lambda)^8. \end{aligned} \tag{1.2}$$

2.d. The codeword $C(x)$ for symbol $x = 00000000$ has length one, and is therefore the shortest codeword. By properties of optimal prefix codes (e.g., see arguments for the optimality of Huffman codes), code C cannot be optimal if the symbol $x = 00000000$ is the least likely symbol, and this occurs provided the condition

$$(1-\lambda)^8 < (1-\lambda)^k\lambda, \quad k = 0, \ldots, 7$$

holds. Since $0 < \lambda < 1$, this is equivalent to $(1-\lambda)^8 < (1-\lambda)^7\lambda$, whence $(1-\lambda) < \lambda$. Thus, the prefix code C defined earlier is not optimal among all prefix codes if $\lambda > \frac{1}{2}$.

3. ———————————————————————————————

3.a. The receiver receives the information encoded in the sequence $\{Y_n,\ n = 01, 2, \ldots\}$ where we have set

$$Y_n = U_n X_n, \quad n = 0, 1, 2, \ldots$$

For each $n = 0, 1, \ldots$, the rv Y_n takes values in the finite alphabet $\{0, 1, \ldots, M\}$, and knowing Y_n is equivalent to knowing both Y_n and U_n – This is so because $Y_n = 0$ if and only if $U_n = 0$ and $X_n \neq 0$.

The entropy rate of the information source as experienced by the receiver is given by

$$\lim_{n\to\infty} \frac{H(Y_0, \ldots, Y_n)}{n+1}$$

provided this limit exists. Fix $n = 0, 1, \ldots$. By the last remark, using the chain rule for entropy we have

$$\begin{aligned} H(Y_0, \ldots, Y_n) &= H(U_0, Y_0, \ldots, U_n, Y_n) \\ &= H(U_0, \ldots, U_n) + H(Y_0, \ldots, Y_n | U_0, \ldots, U_n). \end{aligned} \tag{1.3}$$

For arbitrary $(u_0, u_1, \ldots, u_n)$ in $\{0,1\}^M$ and arbitrary $(y_0, y_1, \ldots, y_n)$ in $\{0, 1, \ldots, M\}$, it is now elementary to see that

$$\begin{aligned} & \mathbb{P}\left[Y_0 = y_0, \ldots, Y_n = y_n | U_0 = u_0, \ldots, U_n = u_n\right] \\ = \ & \mathbb{P}\left[U_0X_0 = y_0, \ldots, U_nX_n = y_n | U_0 = u_0, \ldots, U_n = u_n\right] \\ = \ & \mathbb{P}\left[u_0X_0 = y_0, \ldots, u_nX_n = y_n | U_0 = u_0, \ldots, U_n = u_n\right] \\ = \ & \mathbb{P}\left[u_0X_0 = y_0, \ldots, u_nX_n = y_n\right] \end{aligned} \tag{1.4}$$

since the two sequences $\{X_n,\ n = 0, 1, \ldots\}$ and $\{U_n,\ n = 0, 1, \ldots\}$ are mutually independent. As a result, we find

$$H(Y_0, \ldots, Y_n | U_0 = u_0, \ldots, U_n = u_n) = H(u_0 X_0, \ldots, u_n X_n).$$

If we assume further that the rvs $\{X_n,\ n = 0, 1, \ldots\}$ are i.i.d. rvs, each distributed according to the pmf $\boldsymbol{p}$ on $\mathcal{X}$, then the rvs $u_0 X_0, \ldots, u_n X_n$ are mutually independent, and basic properties of entropy yield

$$\begin{aligned} H(u_0 X_0, \ldots, u_n X_n) &= H(u_0 X_0) + \ldots + H(u_n X_n) \\ &= u_0 H(X_0) + \ldots + u_n H(X_n) \\ &= (u_0 + \ldots + u_n)\, H(X_0). \end{aligned} \tag{1.5}$$

Combining these facts readily leads to

$$H(Y_0, \ldots, Y_n | U_0, \ldots, U_n) = \mathbb{E}\left[U_0 + \ldots + U_n\right] \cdot H(X_0)$$

and we can conclude that

$$H(Y_0, \ldots, Y_n) = H(U_0, \ldots, U_n) + \mathbb{E}\left[U_0 + \ldots + U_n\right] \cdot H(\boldsymbol{p}) \tag{1.6}$$

since $H(X_0) = H(\boldsymbol{p})$.

Finally,

$$\frac{H(Y_0, \ldots, Y_n)}{n+1} = \frac{H(U_0, \ldots, U_n)}{n+1} + \frac{\mathbb{P}\left[U_0 = 1\right] + \ldots + \mathbb{P}\left[U_n = 1\right]}{n+1} \cdot H(\boldsymbol{p})$$

for each $n = 0, 1, \ldots$. Letting n go to infinity, we note that

$$\lim_{n \to \infty} \frac{H(U_0, \ldots, U_n)}{n+1} = H(\mathbb{U})$$

where $H(\mathbb{U})$ is the entropy rate of the sequence $\{U_n,\ n = 0, 1, \ldots\}$ which evolves according to a time-homogeneous irreducible Markov chain with one-step transition probability matrix $\boldsymbol{P}$, say

$$\boldsymbol{P} = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix}.$$

By limiting properties for irreducible Markov chains, the limit

$$a = \lim_{n \to \infty} \frac{\mathbb{P}\left[U_0 = 1\right] + \ldots + \mathbb{P}\left[U_n = 1\right]}{n+1}$$

exists with $0 < a < 1$ satisfying

$$(a, 1-a) = (a, 1-a) \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix}.$$

The conclusion

$$H(\mathbb{U}) + a H(\boldsymbol{p}) \tag{1.7}$$

follows with $a > 0$ as determined above.

3.b. If the rvs $\{U_n,\ n = 0, 1, \ldots\}$ are modeled only as a strictly stationary sequence, then its entropy rate $H(\mathbb{U})$ still exists and is given by

$$\lim_{n\to\infty} \frac{H(U_0, \ldots, U_n)}{n+1} = H(\mathbb{U})$$

and

$$\mathbb{P}[U_0 = 1] + \ldots + \mathbb{P}[U_n = 1] = (n+1)\mathbb{P}[U_0 = 1], \quad n = 0, 1, \ldots$$

Because (1.6) still holds in this case, it is plain that (1.7) will also holds with $a = \mathbb{P}[U_0 = 1]$.

3.c. Assume again that assumptions (i) and (ii) are enforced. The operator at the receiving end forwards to the main office only the symbols which have been correctly received without ever mentioning the possibility that some of the symbols originally transmitted by the source have been corrupted and are missing. We model this situation with the help of the $\{1, \ldots, M\}$-valued rvs $\{Z_k,\ k = 1, 2, \ldots\}$ given by

$$Z_k = X_{\nu_k}, \quad k = 1, 2, \ldots$$

where ν_k is the k^{th} symbol received correctly, namely

$$\nu_{k+1} = \min(n > \nu_k :\ U_n = 1)$$

with

$$\nu_1 = \min(n = 0, 1, \ldots : U_n = 1).$$

For each $k = 1, 2, \ldots$ and arbitrary $x_1, \ldots, x_k$ in $\mathcal{X}$ we have

$$\begin{aligned}
&\mathbb{P}[Z_1 = x_1, \ldots, Z_K = x_k] \\
&= \mathbb{P}[X_{\nu_1} = x_1, \ldots, X_{\nu_K} = x_k] \\
&= \sum_{0\le n_1<\ldots<n_k} \mathbb{P}[X_{\nu_1} = x_1, \ldots, X_{\nu_K} = x_k, \nu_1 = n_1, \ldots, \nu_k = n_k] \\
&= \sum_{0\le n_1<\ldots<n_k} \mathbb{P}[X_{n_1} = x_1, \ldots, X_{n_K} = x_k, \nu_1 = n_1, \ldots, \nu_k = n_k] \\
&= \sum_{0\le n_1<\ldots<n_k} \mathbb{P}[X_{n_1} = x_1, \ldots, X_{n_K} = x_k]\,\mathbb{P}[\nu_1 = n_1, \ldots, \nu_k = n_k] \\
&= \sum_{0\le n_1<\ldots<n_k} \left(\prod_{\ell=1}^{k} \mathbb{P}[X_{n_\ell} = x_\ell]\right) \cdot \mathbb{P}[\nu_1 = n_1, \ldots, \nu_k = n_k] \\
&= \sum_{0\le n_1<\ldots<n_k} \left(\prod_{\ell=1}^{k} p(x_\ell)\right) \cdot \mathbb{P}[\nu_1 = n_1, \ldots, \nu_k = n_k] \\
&= \left(\prod_{\ell=1}^{k} p(x_\ell)\right) \cdot \sum_{0\le n_1<\ldots<n_k} \mathbb{P}[\nu_1 = n_1, \ldots, \nu_k = n_k] \\
&= \left(\prod_{\ell=1}^{k} p(x_\ell)\right) \qquad (1.8)
\end{aligned}$$

as we make use of the fact that the rvs $\{U_n,\ n = 0, 1, \ldots\}$ (hence the rvs $\{\nu_k,\ k = 1, 2, \ldots\}$) are independent of the rvs $\{X_n,\ n = 0, 1, \ldots\}$. It follows that the rvs $\{Z_k,\ k = 1, 2, \ldots\}$ form a sequence of i.i.d. rvs, each distributed according to the pmf $\boldsymbol{p}$. Hence,

$$H(Z_1, \ldots, Z_k) = kH(\boldsymbol{p})$$

and the conclusion

$$\lim_{k\to\infty} \frac{H(Z_1, \ldots, Z_k)}{k} = H(\boldsymbol{p})$$

is obtained.

A careful inspection of the discussion above shows that the conclusion is crucially dependent on assumption (i) (and not on assumption (ii)).

4. _____

4.a. The set $\mathcal{Y}$ of values assumed by Y is the subset of $\mathcal{X}^\star$ given by

$$\mathcal{Y} = \cup_{i=1}^{I} \mathcal{X}^{n_i}.$$

Fix i in $\{1, \ldots, I\}$ and y in $\mathcal{Y}$. Then there exists some j in $\{1, \ldots, I\}$ such that $y = (x_1, \ldots, x_{n_j})$ for some element $(x_1, \ldots, x_{n_j})$ in $\mathcal{X}^{n_j}$. Under the enforced assumptions, we get

$$\begin{aligned}
\mathbb{P}[\nu = n_i, Y = y] &= \mathbb{P}\left[\nu = n_i, \nu = n_j, Y = (x_1, \ldots, x_{n_j})\right] \\
&= \delta_{ij} \cdot \mathbb{P}[\nu = n_j]\, \mathbb{P}\left[(X_1, \ldots, X_{n_j}) = (x_1, \ldots, x_{n_j})\right] \\
&= \delta_{ij} \cdot \mathbb{P}[\nu = n_j] \prod_{k=1}^{n_j} \mathbb{P}[X_k = x_k] \\
&= \delta_{ij} \cdot \mathbb{P}[\nu = n_j] \left(\frac{1}{M}\right)^{n_j}. \qquad (1.9)
\end{aligned}$$

It is now plain for each $i = 1, \ldots, I$ that

$$\mathbb{P}[\nu = n_i, Y = y] = \mathbb{P}[\nu = n_i] \left(\frac{1}{M}\right)^{n_i}, \quad y \in \mathcal{X}^{n_i}$$

and

$$\mathbb{P}[Y = y] = \mathbb{P}[\nu = n_i] \left(\frac{1}{M}\right)^{n_i}, \quad y \in \mathcal{X}^{n_i}.$$

4.b. In order to compute the mutual information $I(\nu; Y)$, we note that

$$I(\nu; Y) = \sum\nolimits_{(n,y)} \mathbb{P}[\nu = n, Y = y] \log_2 \left(\frac{\mathbb{P}[\nu = n, Y = y]}{\mathbb{P}[\nu = n]\, \mathbb{P}[Y = y]}\right)$$

where (n, y) ranges over

$$\cup_{i=1}^{I} \left(\{n_i\} \times \mathcal{X}^{n_i}\right).$$

From earlier calculations, if $n = n_i$ and $y = (x_1, \ldots, x_{n_i})$ for $i = 1, \ldots, I$, then

$$\begin{aligned} \frac{\mathbb{P}[\nu = n_i, Y = y]}{\mathbb{P}[\nu = n_i]\,\mathbb{P}[Y = y]} &= \frac{\mathbb{P}[\nu = n_i]\left(\frac{1}{M}\right)^{n_i}}{\mathbb{P}[\nu = n_i]\,\mathbb{P}[\nu = n_i]\left(\frac{1}{M}\right)^{n_i}} \\ &= \frac{1}{\mathbb{P}[\nu = n_i]}. \end{aligned} \tag{1.10}$$

Consequently,

$$\begin{aligned} I(\nu; Y) &= \sum_{i=1}^{I} \sum_{(x_1,\ldots,x_{n_i}) \in \mathcal{X}^{n_i}} \mathbb{P}[\nu = n_i] \left(\frac{1}{M}\right)^{n_i} \log_2 \left(\frac{1}{\mathbb{P}[\nu = n_i]}\right) \\ &= \sum_{i=1}^{I} \mathbb{P}[\nu = n_i] \log_2 \left(\frac{1}{\mathbb{P}[\nu = n_i]}\right) \end{aligned} \tag{1.11}$$

since $|\mathcal{X}^{n_i}| = M^{n_i}$. As a result,

$$I(\nu; Y) = H(\nu).$$

A more direct derivation of this last fact starts with the decomposition

$$I(\nu; Y) = H(\nu) - H(\nu|Y)$$

and then makes use of the fact $H(\nu|Y) = 0$ because knowledge of Y automatically provides the value of ν – See below.

4.c. The conditional entropy $H(Y|\nu)$ is given by

$$H(Y|\nu) = \sum_{i=1}^{I} \mathbb{P}[\nu = n_i]\, H(Y|\nu = n_i)$$

where for each $i = 1, 2, \ldots, I$, we find

$$H(Y|\nu = n_i) = \log_2 M^{n_i} = n_i \log_2 M$$

since the rv Y, conditionally on $\nu = n_i$, is uniformly distributed on $\mathcal{X}^{n_i}$. Consequently,

$$H(Y|\nu) = \sum_{i=1} I\,(n_i \log_2 M)\,\mathbb{P}[\nu = n_i] = \log_2 M \cdot \mathbb{E}[\nu]\,.$$

4.d. Recall that

$$I(\nu; Y) = H(Y) - H(Y|\nu)$$

whence

$$H(Y) = I(\nu; Y) + H(Y|\nu) = H(\nu) + \log_2 M \cdot \mathbb{E}[\nu]\,.$$

The answer is not too surprising as it formalizes the following decomposition: Upon observing Y, the value of ν is revealed (say, by counting the number of components of Y) and this contributes the term $H(\nu)$ to the uncertainty in Y. Once the value n_i of ν is known, the actual value of Y resolves an uncertainty $n_i \cdot \log_2 M$, which averages to $\log_2 M \cdot \mathbb{E}[\nu]$.