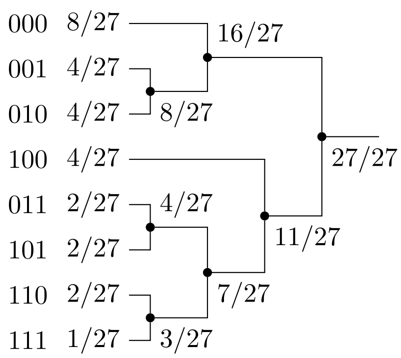
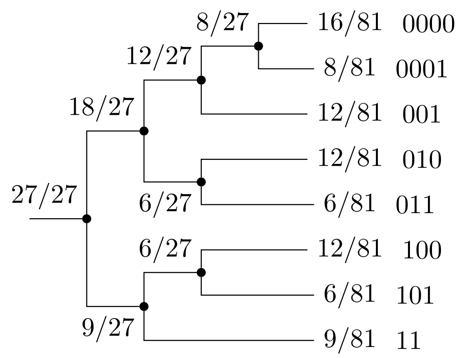
# ECE1502F — Information Theory
## Midterm Test Solution



1. (a) $H(X) = \frac{1}{3}\log_2(3) + \frac{2}{3}\log_2(3/2) \approx 0.9183$ bit/sym

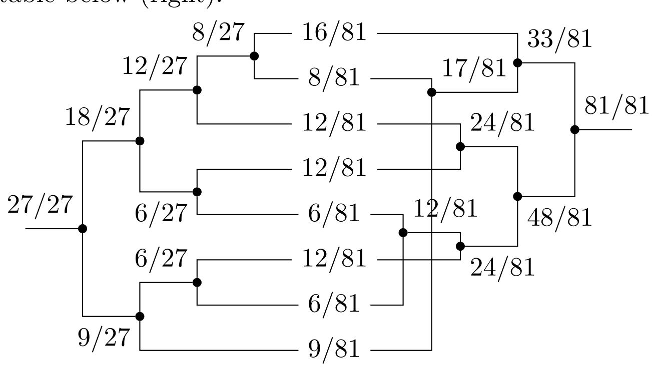
   (b) The probability mass function for the third extension is $\left(\frac{8}{27}, \frac{4}{27}, \frac{4}{27}, \frac{4}{27}, \frac{2}{27}, \frac{2}{27}, \frac{2}{27}, \frac{1}{27}\right)$. Applying the Huffman procedure yields merged symbols with probabilities 3/27, 4/27, 7/27, 8/27, 11/27, 16/27, and 27/27; their sum is the average codeword length, namely $76/27 \approx 2.8148$ bit/triple. One possible Huffman tree is shown below (left). This (fixed-length to variable-length) code achieves a rate of $R = 76/81 \approx 0.938$ bit/sym (2.175% above entropy).




   (c) Applying the Tunstall procedure, starting from a single vertex of probability 27/27 yields the sequence 27/27, 18/27, 12/27, 9/27, 8/27, 6/27, 6/27, of interior node probabilities; their sum is the average phrase length, namely $86/27 \approx 3.185$ sym/phrase. The corresponding tree is shown above (right). This (variable-length to fixed-length) code achieves a rate of $R = 81/86 \approx 0.9419$ bit/sym (2.566% above entropy).

   (d) The phrases corresponding to the Tunstall parsing tree have probability distribution $\left(\frac{16}{81}, \frac{12}{18}, \frac{12}{81}, \frac{12}{81}, \frac{9}{81}, \frac{8}{81}, \frac{6}{81}, \frac{6}{81}\right)$. Applying the Huffman procedure yields merged symbols with probabilities 12/81, 17/81, 24/81, 24/81, 33/81, 48/81, and 81/81; their sum is the average codeword length $239/81 \approx 2.95$ bit/phrase. Thus the fixed length code of length 3 is *not* optimal for this distribution. Combining the Tunstall tree with the Huffman tree yields the (variable-length to variable-length) dual-tree code shown below (left), having rate $R = \frac{239/81 \text{ bit/phrase}}{86/27 \text{ sym/phrase}} = 239/258 \approx 0.9264$ bit/sym (just 0.878% above entropy). The mapping between phrases and binary codewords is given in the table below (right).




| phrase | probability | codeword |
|---|---|---|
| 0000 | 16/81 | 00 |
| 0001 | 8/81 | 010 |
| 001 | 12/81 | 100 |
| 010 | 12/81 | 101 |
| 011 | 6/81 | 1110 |
| 100 | 12/81 | 110 |
| 101 | 6/81 | 1111 |
| 11 | 9/81 | 011 |

2. Note that by setting $\epsilon = \delta H(X)$, we get that $A_\epsilon^{(n)} = B_\delta^{(n)}$. The properties to be proved then essentially follow immediately from Theorem 3.1.2 in Cover and Thomas (C&T). However, since we want to have $1-\delta$ (not $1-\epsilon$) as a lower bound on $\Pr(B_\delta^{(n)})$, we rewrite the proof of C&T Theorem 3.1.2 as follows.

(a) In the following, $\Leftrightarrow$ means "if and only if" or "is equivalent to". By definition we have

$$\begin{aligned}(x_1,\ldots,x_n)\in B_\delta^{(n)} &\Leftrightarrow \left|-\frac{1}{n}\log p(x_1,\ldots,x_n)-H(X)\right|\le \delta H(X)\\ &\Leftrightarrow -\delta H(X)\le -\frac{1}{n}\log p(x_1,\ldots,x_n)-H(X)\le \delta H(X)\\ &\Leftrightarrow H(X)-\delta H(X)\le -\frac{1}{n}\log p(x_1,\ldots,x_n)\le H(X)+\delta H(X)\\ &\Leftrightarrow -nH(X)(1-\delta)\ge \log p(x_1,\ldots,x_n)\ge -nH(X)(1+\delta)\\ &\Leftrightarrow 2^{-nH(X)(1+\delta)}\le p(x_1,\ldots,x_n)\le 2^{-nH(X)(1-\delta)}\end{aligned}$$

(b) Let $X_1, X_2, \ldots$ be i.i.d., where $H(X_1)=H(X)$. By the AEP, $-\frac{1}{n}\log p(X_1,\ldots,X_n)\xrightarrow{p} H(X)$, i.e., for any $\epsilon>0$ and any $\gamma>0$, there exists a positive integer $n_0$ such that the inequality

$$\Pr\left(\left|-\frac{1}{n}\log p(X_1,\ldots,X_n)-H(X)\right|<\epsilon\right)\ge 1-\gamma$$

holds for all $n\ge n_0$. Setting $\epsilon=\delta H(X)$ and $\gamma=\delta$ and recognizing that the enclosed event is $B_\delta^{(n)}$, we see that $\Pr(B_\delta^{(n)})\ge 1-\delta$ when $n$ is sufficiently large.

(c) We have

$$1\ge \Pr(B_\delta^{(n)})=\sum_{(x_1,\ldots,x_n)\in B_\delta^{(n)}} p(x_1,\ldots,x_n)\ge \sum_{(x_1,\ldots,x_n)\in B_\delta^{(n)}} 2^{-n(1+\delta)H(X)}=\left|B_\delta^{(n)}\right|2^{-n(1+\delta)H(X)},$$

where the second inequality follows from part (a), which implies that $\left|B_\delta^{(n)}\right|\le 2^{n(1+\delta)H(X)}$.

(d) From (b) we have, when $n$ is sufficiently large, that

$$1-\delta\le \Pr(B_\delta^{(n)})=\sum_{(x_1,\ldots,x_n)\in B_\delta^{(n)}} p(x_1,\ldots,x_n)\le \sum_{(x_1,\ldots,x_n)\in B_\delta^{(n)}} 2^{-n(1-\delta)H(X)}=\left|B_\delta^{(n)}\right|2^{-n(1-\delta)H(X)},$$

where the second inequality follows from part (a), which implies that $\left|B_\delta^{(n)}\right|\ge (1-\delta)2^{n(1-\delta)H(X)}$.

(e) For $\epsilon_1<\epsilon_2$ the containment $A_{\epsilon_1}^{(n)}\subseteq A_{\epsilon_2}^{(n)}$ holds, i.e., the typical set increases as $\epsilon$ increases. When $\epsilon=\delta H(X)$ we have

$$A_\epsilon^{(n)}=\left\{(x_1,\ldots,x_n)\in\mathcal{X}^n:\left|-\frac{1}{n}\log p(x_1,\ldots,x_n)-H(X)\right|\le\delta H(X)\right\}=B_\delta^{(n)}.$$

When $\epsilon<\delta H(X)$ then $A_\epsilon^{(n)}\subseteq B_\delta^{(n)}$ and when $\epsilon>\delta H(X)$ then $B_\delta^{(n)}\subseteq A_\epsilon^{(n)}$.

3. (a) For $i\in\mathcal{X}$, let $p_i=P(X=i)$. We want to show that $E(X)-H(X)\ge 0$. To this end, we compute

$$\begin{aligned}E(X)-H(X)&=\sum_{i\ge1} ip_i+\sum_{i\ge1}p_i\log_2 p_i=-\sum_{i\ge1}p_i\log_2\left(2^{-i}\right)+\sum_{i\ge1}p_i\log_2 p_i=\sum_{i\ge1}p_i\log_2\frac{p_i}{2^{-i}}\\ &\ge\left(\sum_{i\ge1}p_i\right)\log_2\frac{\sum_{i\ge1}p_i}{\sum_{i\ge1}2^{-i}}=-\log_2\left(\sum_{i\ge1}2^{-i}\right)=0,\end{aligned}$$

where the inequality follows from the log sum inequality and the last equality follows from the fact that $\sum_{i\ge1}2^{-i}=1$. This latter fact also means $E(X)-H(X)$ is equal to the relative entropy between the given distribution and the geometric distribution with parameter $1/2$.

We may also reason as follows. Consider the infinite binary prefix code $\{0,10,110,1110,\ldots\}$ having one codeword of each positive integer length. Assigning the codeword of length $i$ to outcome $i$ gives an average codeword length $L=\sum_{i\ge1}ip_i=E(X)$. Since this code is uniquely decodable, we must have $L\ge H(X)$, from which we deduce that $E(X)\ge H(X)$.

(b) Equality is achieved when $p_i = 2^{-i}$, i.e., when $X$ has a geometric distribution.

(c) If $\mathcal{Y} = \{0, 1, 2, \ldots\}$, let $X = 1 + Y$, so $E(X) = 1 + E(Y)$. Since $X$ is a one-to-one function of $Y$, $H(X) = H(Y)$. From (a), it follows that

$$H(Y) = H(X) \leq E(X) = 1 + E(Y).$$

(d) Equality is achieved if and only if $P(X = i) = P(Y = i-1) = 2^{-i}$, $i \geq 1$, i.e., $P(Y = i) = 2^{-(i+1)}$, $i \geq 0$.

4. (a) Let $p_Y^*(y) = \sum_{x\in\mathcal{X}} p_X^*(x) p_{Y|X}(y \mid x)$ be the distribution over $\mathcal{Y}$ induced by $p_X^*$, and define $p_Y'(y)$ and $p_Y''(y)$ similarly. Then

$$\begin{aligned}
&I(p_X^*; p_{Y|X}) - \lambda I(p_X'; p_{Y|X}) - (1-\lambda) I(p_X''; p_{Y|X}) \\
&\qquad = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p^*(x) p(y \mid x) \log \frac{p^*(x)p(y \mid x)}{p^*(x)p^*(y)} - \lambda \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p'(x) p(y \mid x) \log \frac{p'(x)p(y \mid x)}{p'(x)p'(y)} \\
&\qquad\quad - (1-\lambda) \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p''(x) p(y \mid x) \log \frac{p''(x)p(y \mid x)}{p''(x)p''(y)} \\
&\qquad = \lambda \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p'(x) p(y \mid x) \log \frac{p'(y)}{p^*(y)} + (1-\lambda) \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p''(x) p(y \mid x) \log \frac{p''(y)}{p^*(y)} \\
&\qquad = \lambda \sum_{y\in\mathcal{Y}} p'(y) \log \frac{p'(y)}{p^*(y)} + (1-\lambda) \sum_{y\in\mathcal{Y}} p''(y) \log \frac{p''(y)}{p^*(y)} \\
&\qquad = \lambda D(p_Y' \| p_Y^*) + (1-\lambda) D(p_Y'' \| p_Y^*) \\
&\qquad \geq 0.
\end{aligned}$$

We may also reason as follows. Let $S \in \{0, 1\}$ be a Bernoulli random variable, with $P(S = 0) = \lambda$ and $P(S = 1) = 1 - \lambda$. Define a random variable $X$ over $\mathcal{X}$ (dependent upon $S$) so that

$$\begin{aligned}
p_{X|S}(x \mid S = 0) &= p_X'(x) \\
p_{X|S}(x \mid S = 1) &= p_X''(x)
\end{aligned}$$

We think of $S$ as a "distribution selector" for $X$. The marginal distribution for $X$ is then

$$p_X(x) = \sum_{s=0}^{1} p_{X|S}(x \mid S = s) p_S(s) = \lambda p_X'(x) + (1-\lambda) p_X''(x).$$

Define $Y$ so that $S \to X \to Y$ forms a Markov chain, with $p_{Y|X}(y \mid x)$ fixed. From the chain rule for mutual information we deduce that

$$I(X, S; Y) = I(X; Y) + I(S; Y \mid X) = I(S; Y) + I(X; Y \mid S).$$

Now, since $S \to X \to Y$ is a Markov chain, we have $I(S; Y \mid X) = 0$. Since $I(S; Y) \geq 0$, we find that

$$I(X; Y) \geq I(X; Y \mid S) = \lambda I(X; Y \mid S = 0) + (1-\lambda) I(X; Y \mid S = 1).$$

Changing notation, this latter inequality can be written as

$$I(\lambda p_X' + (1-\lambda) p_X''; p_{Y|X}) \geq \lambda I(p_X'; p_{Y|X}) + (1-\lambda) I(p_X''; p_{Y|X}),$$

which shows that $I(p_X; p_{Y|X})$ is concave in $p_X$ when $p_{Y|X}$ is fixed.

(b) Let $p^*_{X|Y}(x \mid y) = p(x)p^*(y \mid x)/p^*(y)$, and define $p'_{X|Y}$ and $p''_{X|Y}$ similarly. Then

$$
\begin{aligned}
&\lambda I(p_X; p'_{Y|X}) + (1-\lambda) I(p_X; p''_{Y|X}) - I(p_X; p^*_{Y|X}) \\
&\qquad = \lambda \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x)p'(y \mid x) \log \frac{p'(y)p'(x \mid y)}{p'(y)p(x)} + (1-\lambda) \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x)p''(y \mid x) \log \frac{p''(y)p''(x \mid y)}{p''(y)p(x)} \\
&\qquad\quad - \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x)p^*(y \mid x) \log \frac{p^*(y)p^*(x \mid y)}{p^*(y)p(x)} \\
&\qquad = \lambda \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x)p'(y \mid x) \log \frac{p'(x \mid y)}{p^*(x \mid y)} + (1-\lambda) \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x)p''(y \mid x) \log \frac{p''(x \mid y)}{p^*(x \mid y)} \\
&\qquad = \lambda \sum_{y\in\mathcal{Y}} p'(y) \sum_{x\in\mathcal{X}} p'(x \mid y) \log \frac{p'(x \mid y)}{p^*(x \mid y)} + (1-\lambda) \sum_{y\in\mathcal{Y}} p''(y) \sum_{x\in\mathcal{X}} p''(x \mid y) \log \frac{p''(x \mid y)}{p^*(x \mid y)} \\
&\qquad = \lambda \sum_{y\in\mathcal{Y}} p'(y) D\left(p'_{X|Y=y} \| p^*_{X|Y=y}\right) + (1-\lambda) \sum_{y\in\mathcal{Y}} p''(y) D\left(p''_{X|Y=y} \| p^*_{X|Y=y}\right) \\
&\qquad \geq 0
\end{aligned}
$$

We may also reason as follows. Let $S \in \{0, 1\}$ be a Bernoulli random variable, independent of $X$, with $P(S = 0) = \lambda$ and $P(S = 1) = 1 - \lambda$. Define $Y$ (dependent upon both $X$ and $S$) so that

$$
\begin{aligned}
p_{Y|X,S}(y \mid x, S = 0) &= p'_{X|Y}(y \mid x) \\
p_{Y|X,S}(y \mid x, S = 1) &= p''_{X|Y}(y \mid x).
\end{aligned}
$$

We think of $S$ as a "channel selector" relating $X$ with $Y$. The conditional probability mass function for $Y$ given $X$ is then

$$
p_{Y|X}(y \mid x) = \sum_{s=0}^{1} p_{Y|X,S}(y \mid x, s) p_S(s) = \lambda p'_{Y|X}(y \mid x) + (1-\lambda) p''_{Y|X}(y \mid x).
$$

From the chain rule for mutual information we deduce that

$$
I(X; S, Y) = I(X; Y) + I(X; S \mid Y) = I(X; S) + I(X; Y \mid S).
$$

Since $X$ and $S$ are independent, we have $I(X; S) = 0$. Since $I(X; S \mid Y) \geq 0$, we find that

$$
I(X; Y) \leq I(X; Y \mid S) = \lambda I(X; Y \mid S = 0) + (1-\lambda) I(X; Y \mid S = 1).
$$

Changing notation, this latter inequality can be written as

$$
I(p_X; \lambda p'_{Y|X} + (1-\lambda) p''_{Y|X}) \leq \lambda I(p_X; p'_{Y|X}) + (1-\lambda) I(p_X; p''_{Y|X}),
$$

which shows that $I(p_X; p_{Y|X})$ is convex in $p_{Y|X}$ when $p_X$ is fixed.

*Remark:* This problem corresponds to Theorem 2.7.4 in Cover and Thomas, where yet another proof is given.

5. (a) The joint probability mass function for $X$ and $Y$ is given as

$$
p_{X,Y}(x, y) = \begin{cases} \frac{p}{2}, & \text{if } x = y; \\ \frac{1-p}{2}, & \text{if } x \neq y. \end{cases}
$$

Both $X$ and $Y$ are uniformly distributed. Thus

$$
\begin{aligned}
I(X; Y) &= \sum_x \sum_y p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \\
&= p \log(2p) + (1-p) \log(2(1-p)) \\
&= 1 - \mathcal{H}(p).
\end{aligned}
$$

(b) If the player has capital $C_{n-1}$ at time $n-1$, then their capital at time $n$ is

$$\begin{aligned} C_n &= \begin{cases} C_{n-1}(1+q) & \text{if the guess is correct,} \\ C_{n-1}(1-q) & \text{if the guess is incorrect} \end{cases} \\ &= C_{n-1}(1+q)^{Z_n}(1-q)^{(1-Z_n)}. \end{aligned}$$

Applying this recursively yields the formula given. The growth rate,

$$\begin{aligned} R_N &= \frac{1}{N}\log_2\frac{C_N}{C_0} \\ &= \frac{1}{N}\log_2\prod_{n=1}^{N}(1+q)^{Z_n}(1-q)^{1-Z_n} \\ &= \frac{1}{N}\sum_{n=1}^{N} Z_n\log_2(1+q) + (1-Z_n)\log_2(1-q) \\ &= \log_2(1-q) + \log_2\left(\frac{1+q}{1-q}\right)\left(\frac{1}{N}\sum_{n=1}^{N} Z_n\right) \end{aligned}$$

The expected growth rate is therefore

$$\begin{aligned} E[R_N] &= \log_2(1-q) + \log_2\left(\frac{1+q}{1-q}\right)\left(\frac{1}{N}\sum_{n=1}^{N} E[Z_n]\right) \\ &= \log_2(1-q) + p\log_2\left(\frac{1+q}{1-q}\right) \\ &= p\log_2(1+q) + (1-p)\log_2(1-q) \\ &= -\left(p\log_2\frac{p}{1+q} + (1-p)\log_2\frac{1-p}{1-q}\right) + p\log_2 p + (1-p)\log_2(1-p) \\ &\le -(p+1-p)\log_2\frac{p+1-p}{1+q+1-q} - \mathcal{H}(p) \\ &= 1-\mathcal{H}(p), \end{aligned}$$

where the inequality follows from the log sum inequality. Equality is achieved when $p/(1+q) = (1-p)/(1-q)$, i.e., when $q = 2p-1$, in which case $E[R_N] = 1-\mathcal{H}(p) = I(X;Y)$. Thus, the more information the player has about the coin toss, the greater will be the expected growth rate.

Now, let $Y_n = (1+q)^{Z_n}(1-q)^{1-Z_n}$. Then $C_N = C_0\prod_{n=1}^{N} Y_n$. Since the $Z_n$ are independent and identically distributed, so are the $Y_n$ and hence

$$E[C_N] = C_0(E[Y_1])^N.$$

Now, $E[Y_1] = p(1+q) + (1-p)(1-q) = 1 + q(2p-1)$. Since $p > 1/2$, $2p-1 > 0$ and $E[Y_1] > 1$, so $E[C_N] = C_0(1+q(2p-1))^N$ is maximized by setting $q = 1$.

(c) I would use $q = 2p-1$, since my wealth would almost surely grow at the maximum exponential rate. On the other, choosing $q = 1$, which maximizes my expected wealth, would not be a good strategy, since the expected value would not typically be achieved. Indeed, I would almost surely guess wrong and lose everything. Note that $C_N$ is a *product* of (functions of) random variables, whereas $R_N$ is a *sum* of (functions of) random variables. The weak law of large numbers applies to the latter, but *not* to the former.