University of Toronto
December 14, 2021

ECE1502F — Information Theory
Final Exam Solution

Department of Electrical
& Computer Engineering

1. (*A Markov Source*)

(a) The probability transition matrix is

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1/2 & 0 & 1/4 & 1/4 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1/4 & 1/4 & 1/2 \\ 0 & 0 & 0 & 1/2 & 1/2 \end{bmatrix}$$

where the rows and columns are arranged in the order (a, b, c, d, e) and the element in row x and column y represents the probability that the next state is y given that the present state is x. This matrix is doubly stochastic, and so by C&T Problem 4.1, for $\mu = (1/5, 1/5, 1/5, 1/5, 1/5)$ we have $\mu P = \mu$. Thus the stationary distribution is uniform.

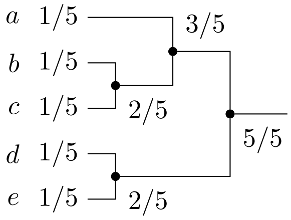
(b) Let X_1 denote the present state and X_2 the next state. Then

$$\begin{aligned} H(X_2 \mid X_1 = a) &= H(0, 1, 0, 0, 0) = 0, \\ H(X_2 \mid X_1 = b) &= H(1/2, 0, 1/4, 1/4, 0) = 3/2 \text{ bit}, \\ H(X_2 \mid X_1 = c) &= H(1/2, 0, 1/2, 0, 0) = 1 \text{ bit}, \\ H(X_2 \mid X_1 = d) &= H(0, 0, 1/4, 1/4, 1/2) = 3/2 \text{ bit}, \\ H(X_2 \mid X_1 = e) &= H(0, 0, 0, 1/2, 1/2) = 1 \text{ bit}. \end{aligned}$$

In steady state, $H(X_2 \mid X_1) = \frac{1}{5}(0 + 3/2 + 1 + 3/2 + 1) = 1$ bit.

(c) The encoding of X_1 doesn't need to be particularly efficient, since it is only done once at the start of the encoding process. For example a fixed-length code (of 3 bits) could be chosen. Nevertheless we will encode X_1 according to the Huffman code indicated in the following table (the corresponding to tree is also shown).

$X_1 =$	a	b	c	d	e
Codeword	00	010	011	10	11



We will encode symbols *conditionally*. Given that given X_i, we encode X_{i+1} according to the following table.

	$X_{i+1} = a$	$X_{i+1} = b$	$X_{i+1} = c$	$X_{i+1} = d$	$X_{i+1} = e$
$X_i = a$	—	ϵ	—	—	—
$X_i = b$	0	—	10	11	—
$X_i = c$	0	—	1	—	—
$X_i = d$	—	—	10	11	0
$X_i = e$	—	—	—	0	1

The transitions labeled '—' do not occur. If $X_i = a$, we do not output a symbol, but instead, implicitly transfer to state $X_{i+1} = b$. (The symbol ϵ in the table denotes the 'empty string'.) The codes in each row of the table are optimal prefix codes for the corresponding dyadic distribution. We have not bothered to draw the corresponding binary trees.

If the input terminates in state a, we output a '1', followed by end-of-file, which is not a 'valid' output sequence from state b. This 'exception' condition can be used by the decoder to suppress the 'b' that usually follows the occurence of an 'a'. Thus, an input consisting of a single 'a' produces the output string 001, while an input consisting of the input ab produces the output string 00.

This encoding scheme is efficient because to encode n input symbols requires, on average, (assuming the initial state is chosen uniformly at random) 2.4 bits to encode the starting state, plus $n-1$ bits to encode the succeeding symbols, plus at most 1 exception bit. Thus to encode n input symbols requires, on average, at most $2.4+n$ output bits, or $1+2.4/n$ output bits per symbol. For large n, this number converges to the entropy rate of 1 bit per symbol.

(d) The input string *bcababdedcc* produces the output string

$$01010001100101 = \underbrace{010}_{b}\underbrace{10}_{c}\underbrace{0}_{ab}\underbrace{0}_{ab}\underbrace{11}_{d}\underbrace{0}_{e}\underbrace{0}_{d}\underbrace{10}_{c}\underbrace{1}_{c}.$$

2. (*Maximum Entropy*)

(a) Let p_i denote $P(X=i)$ and let $q_i = \frac{1}{m+1}\left(\frac{m}{m+1}\right)^i$. We have

$$\begin{aligned}
0 &\le D(p||q) \\
&= \sum_{i\ge 0} p_i \log \frac{p_i}{q_i} \\
&= \sum_{i\ge 0} p_i \log p_i - \sum_{i\ge 0} p_i \log q_i \\
&= -H(X) - \sum_{i\ge 0} p_i \log\left(\frac{1}{m+1}\left(\frac{m}{m+1}\right)^i\right) \\
&= -H(X) + \sum_{i\ge 0} p_i \log(m+1) - \sum_{i\ge 0} ip_i \log(m) + \sum_{i\ge 0} ip_i \log(m+1) \\
&= -H(X) + \log(m+1) - m\log(m) + m\log(m+1) \\
&= -H(X) + (m+1)\log(m+1) - m\log(m),
\end{aligned}$$

from which we deduce that $H(X) \le (m+1)\log(m+1) - m\log(m)$. Equality is achieved if and only if $D(p||q)=0$ if and only if $p_i = q_i$ for all i.

(b) Let $f(x)$ be an arbitrary probability density function supported on $[a,b]$, and let $u(x) = 1/(b-a)$ be the uniform distribution over $[a,b]$. We then have

$$\begin{aligned}
0 &\le D(f||u) \\
&= \int_a^b f(x)\log\frac{f(x)}{u(x)}\,\mathrm{d}x \\
&= \int_a^b f(x)\log f(x)\,\mathrm{d}x - \int_a^b f(x)\log u(x)\,\mathrm{d}x \\
&= -h(X) + \int_a^b f(x)\log(b-a)\,\mathrm{d}x \\
&= -h(X) + \log(b-a),
\end{aligned}$$

from which we deduce that $h(X) \le \log(b-a)$. Equality is achieved if and only if $f(x) = u(x)$, so we see that the uniform distribution has maximum entropy.

(c) We have $h(X_1) = \log(b-a)$. Furthermore, we have

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \prod_{i=1}^n f_{X_i}(x_i) = \begin{cases} (b-a)^{-n}, & x_1 \in [a,b],\ldots,x_n \in [a,b], \\ 0, & \text{otherwise.} \end{cases}$$

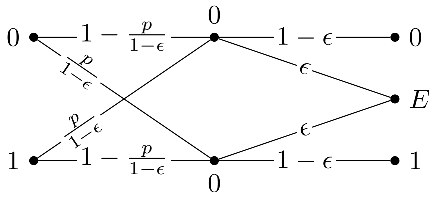
and thus $\frac{-1}{n}\log f(x_1,\ldots,x_n) = \log(b-a) = h(X_1)$ for all $(x_1,\ldots,x_n) \in [a,b]^n$. In other words, irrespective of ϵ, *every* vector in $[a,b]^n$ is typical. The typical set is a hypercube.

3. (*Errors-and-Erasures Channel*)

   (a) The errors-and-erasures channel can be obtained as the cascade of a binary symmetric channel with crossover probability $p/(1-\epsilon)$ and a binary erasure channel with erasure probability ϵ, as shown in the figure below. Then by the result of C&T Problem 7.27 we have that

$$C = (1-\epsilon)\left(1 - \mathcal{H}\left(\frac{p}{1-\epsilon}\right)\right),$$

   where $\mathcal{H}$ denotes the binary entropy function.



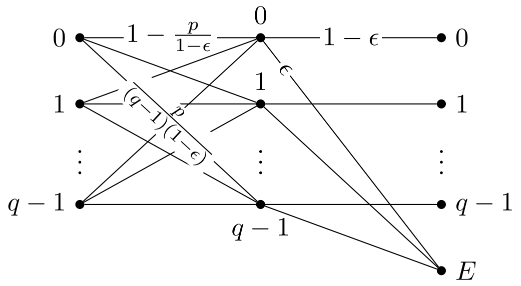
   (b) When $p = 0$, we recover $C = (1-\epsilon)(1-\mathcal{H}(0)) = 1-\epsilon$ (as we must).

   (c) When $\epsilon = 0$, we recover $c = (1-0)(1-\mathcal{H}(p)) = 1-\mathcal{H}(p)$ (as we must).

   (d) Let C be a code designed for the binary symmetric channel with crossover probability $p/(1-\epsilon)$. Send the symbols of a codeword one-by-one. Whenever the channel output is E, the receiver asks for a retransmission; otherwise it passes the received symbol to a decoder for C. The decoder for C effectively sees a sequence of received symbols from a binary symmetric channel without any erasures at all. If C has a rate that approaches $1-\mathcal{H}(p/(1-\epsilon))$, the overall rate of transmission approaches $(1-\epsilon)(1-\mathcal{H}(p/(1-\epsilon)))$.

   (e) The q-ary erasure channel is the cascade of a q-ary symmetric channel (with probability of error $p/(1-\epsilon)$) with a q-ary erasure channel, as shown in the figure below. We can model a q-ary symmetric channel with error probability α as $Y = X + Z \bmod q$, where Z is independent of X with $P(Z=0) = 1-\alpha$, and $P(Z=z) = \alpha/(q-1)$ when $z \neq 0$. The capacity of this channel is achieved by a uniform input distribution, and is given as $\log(q) - H(Z)$. Then by the result of C&T Problem 7.27 we have that

$$C = (1-\epsilon)\left(\log(q) - H\left(1 - \frac{p}{1-\epsilon}, \frac{p}{(q-1)(1-\epsilon)}, \ldots, \frac{p}{(q-1)(1-\epsilon)}\right)\right).$$



4. (*Channel Reduction*)

   (a) Let $X \in \mathcal{X}$ denote the channel input, Y the channel output (before reduction) and Z the output of the reduced channel. Let y_1 and y_2 be output letters combined in the reduced channel, and let z be the corresponding output of the reduced channel. We assume that for all $x \in \mathcal{X}$ and for some constant k,

$$P(Y = y_1 \mid X = x) = kP(Y = y_2 \mid X = x),$$

   which is equivalent to saying that the corresponding columns in the channel matrix are proportional. Fix any channel input distribution $p(x)$.

First note that $p(y_2) = \sum_{x \in \mathcal{X}} p(y_2 \mid x)p(x) = \sum_{x \in \mathcal{X}} kp(y_1 \mid x)p(x) = kp(y_1)$. It follows that $p(z) = p(y_1) + p(y_2) = (1+k)p(y_1)$.

Furthermore, for any $x \in \mathcal{X}$,

$$p(x \mid y_2) = \frac{p(y_2 \mid x)p(x)}{p(y_2)} = \frac{kp(y_1 \mid x)p(x)}{kp(y_1)} = p(x \mid y_1),$$

and

$$p(x \mid z) = \frac{p(z \mid x)p(x)}{p(z)} = \frac{((y_1 \mid x) + p(y_2 \mid x))p(x)}{(1+k)p(y_1)} = \frac{(1+k)p(y_1 \mid x)p(x)}{(1+k)p(y_1)} = p(x \mid y_1).$$

Since $p(x \mid y_1) = p(x \mid y_2) = p(x \mid z)$ for all $x \in \mathcal{X}$, we have $H(X \mid Y = y_1) = H(X \mid Y = y_2) = H(X \mid Z = z)$.

Finally note that

$$\begin{aligned} I(X;Z) - I(X;Y) &= H(X \mid Y) - H(X \mid Z) \\ &= H(X \mid y_1)p(y_1) + H(X \mid y_2)p(y_2) - H(X \mid Z)p(z) \\ &= H(X \mid y_1)(1+k)p(y_1) - H(X \mid y_1)(1+k)p(y_1) \\ &= 0; \end{aligned}$$

thus $I(X;Y) = I(X;Z)$. Since the original channel and the reduced channel have exactly the same mutual information between channel input and channel output for every input distribution, they have the same capacity.

(b) Let $C(M)$ denote the capacity of a channel with channel matrix M. Applying the reduction rule, we find that

$$C\left(\begin{bmatrix} \frac{5}{32} & \frac{3}{8} & \frac{5}{32} & \frac{5}{16} \\ \frac{7}{32} & \frac{1}{8} & \frac{7}{32} & \frac{7}{16} \end{bmatrix}\right) = C\left(\begin{bmatrix} \frac{5}{16} & \frac{3}{8} & \frac{5}{16} \\ \frac{7}{16} & \frac{1}{8} & \frac{7}{16} \end{bmatrix}\right) = C\left(\begin{bmatrix} \frac{5}{8} & \frac{3}{8} \\ \frac{7}{8} & \frac{1}{8} \end{bmatrix}\right),$$

which is the channel law for a binary asymmetric channel.

Consider a binary asymmetric channel BAC(p,q) with input alphabet $\mathcal{X} = \{0,1\}$, output alphabet $\mathcal{Y} = \{0,1\}$, and channel matrix

$$\begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}$$

Let r denote $P(X=0)$. The mutual information between channel input and output, expressed as a function of r, is then

$$\begin{aligned} I(r) &= H(Y) - H(Y \mid X) \\ &= \mathcal{H}((1-p)r + q(1-r)) - r\mathcal{H}(p) - (1-r)\mathcal{H}(q) \\ &= \mathcal{H}(r(1-p-q)+q) - r\mathcal{H}(p) - (1-r)\mathcal{H}(q) \end{aligned}$$

When $p+q=1$, we see that $I(r)=0$ for any r; thus we will assume that $p+q \neq 1$. We must choose r to maximize $I(r)$. We have

$$\frac{\mathrm{d}}{\mathrm{d}r}I(r) = \log\left(\frac{1-r(1-p-q)-q}{r(1-p-q)+q}\right)(1-p-q) - \mathcal{H}(p) + \mathcal{H}(q),$$

where the base of the logarithm must match the base of the logarithm used to define the binary entropy function $\mathcal{H}$. We will use logarithms to base two. The derivative is zero when

$$\frac{1-r(1-p-q)-q}{r(1-p-q)+q} = Z(p,q)$$

where

$$Z(p,q) = 2^{\frac{\mathcal{H}(p)-\mathcal{H}(q)}{1-p-q}}.$$

Solving for r we get

$$r^* = P^*(X=0) = \frac{1-q-qZ(p,q)}{(1-p-q)(1+Z(p,q))},$$

where the $*$ denotes that this is the capacity-achieving parameter. Substituting r^* into I, we get

$$C(p,q) = \mathcal{H}\left(\frac{1}{1+Z(p,q)}\right) - \frac{1-q-qZ(p,q)}{(1-p-q)(1+Z(p,q))}\mathcal{H}(p) - \frac{Z(p,q)-p-pZ(p,q)}{(1-p-q)(1+Z(p,q))}\mathcal{H}(q)$$

This is already a nice expression, but we can simplify it further by noting that

$$\begin{aligned}
\mathcal{H}\left(\frac{1}{1+Z(p,q)}\right) &= -\frac{1}{1+Z(p,q)}\log_2\left(\frac{1}{1+Z(p,q)}\right) - \frac{Z(p,q)}{1+Z(p,q)}\log_2\left(\frac{Z(p,q)}{1+Z(p,q)}\right) \\
&= \log_2(1+Z(p,q)) - \frac{Z(p,q)}{1+Z(p,q)}\log_2(Z(p,q)) \\
&= \log_2(1+Z(p,q)) - \frac{Z(p,q)\mathcal{H}(p) - Z(p,q)\mathcal{H}(q)}{(1-p-q)(1+Z(p,q))}
\end{aligned}$$

Substituting this into the expression for $C(p,q)$ (and still assuming that $p+q \neq 1$), we get

$$\begin{aligned}
C(p,q) &= \log_2(1+Z(p,q)) - \frac{Z(p,q)\mathcal{H}(p) - Z(p,q)\mathcal{H}(q)}{(1-p-q)(1+Z(p,q))} \\
&\quad - \frac{1-q-qZ(p,q)}{(1-p-q)(1+Z(p,q))}\mathcal{H}(p) - \frac{Z(p,q)-p-pZ(p,q)}{(1-p-q)(1+Z(p,q))}\mathcal{H}(q) \\
&= \log_2(1+Z(p,q)) - \frac{1-q-qZ(p,q)+Z(p,q)}{(1-p-q)(1+Z(p,q))}\mathcal{H}(p) - \frac{Z(p,q)-p-pZ(p,q)-Z(p,q)}{(1-p-q)(1+Z(p,q))}\mathcal{H}(q) \\
&= \log_2(1+Z(p,q)) - \frac{(1-q)(1+Z(p,q))}{(1-p-q)(1+Z(p,q))}\mathcal{H}(p) + \frac{p(1+Z(p,q))}{(1-p-q)(1+Z(p,q))}\mathcal{H}(q) \\
&= \log_2(1+Z(p,q)) - \frac{1-q}{1-p-q}\mathcal{H}(p) + \frac{p}{1-p-q}\mathcal{H}(q)
\end{aligned}$$

To summarize, the capacity of the binary asymmetric channel with crossover parameters p and q is given as

$$C(p,q) = \begin{cases} \log_2(1+Z(p,q)) - \frac{1-q}{1-p-q}\mathcal{H}(p) + \frac{p}{1-p-q}\mathcal{H}(q), & \text{when } p+q \neq 1; \\ 0, & \text{when } p+q = 1, \end{cases}$$

$$\text{where } Z(p,q) = 2^{\frac{\mathcal{H}(p)-\mathcal{H}(q)}{1-p-q}},$$

$$\text{achieved, when } p+q \neq 1, \text{ by setting } P(X=0) = \frac{1-q-qZ(p,q)}{(1-p-q)(1+Z(p,q))}.$$

In our case $p = 3/8$ and $q = 7/8$. We obtain $r^* \approx 0.4699$, and $C \approx 0.0625$ bit/channel use.

5. (*Composite Channels*)

(a) The channel transition matrix for this channel is

$$M = \begin{bmatrix} 1-p & p \\ p & 1-p \end{bmatrix}\begin{bmatrix} 1-p & p \\ p & 1-p \end{bmatrix} = \begin{bmatrix} 1-2p(1-p) & 2p(1-p) \\ 2p(1-p) & 1-2p(1-p) \end{bmatrix}$$

This is a binary symmetric channel with crossover probability $2p(1-p)$, hence

$$C_{\text{cascade}} = 1 - \mathcal{H}(2p(1-p)).$$

(b) Clearly $C_{\text{recode}} \leq C_{\text{BSC}}(p) = 1 - \mathcal{H}(p)$, since the bit rate through the second channel is limited by this amount. Intuitively, since the recoder can decode/code reliably at all rates $R < C_{\text{BSC}}(p)$, we expect $C_{\text{recode}} = C_{\text{BSC}}(p)$.

(c) To prove that all rates $R < C_{\text{BSC}}(p)$ are indeed achievable, fix a rate $R < C_{\text{BSC}}$ and $\epsilon > 0$. By definition, for n sufficiently large, there is some $(2^{nR}, n)$ code C^* for the BSC with maximal probability of error $\lambda^{(n)} \leq \epsilon/2$. We assume that the transmitter, recoder, and receiver all operate with the code C^*. For any message $w \in \{1, 2, \ldots, 2^{nR}\}$, the probability of correct decoding at the receiver is at least as large as the probability that the recoder and the receiver both decode correctly, i.e.,

$$1 - \lambda_w \geq (1 - \epsilon/2) \cdot (1 - \epsilon/2) = 1 - \epsilon + \epsilon^2/4 \geq 1 - \epsilon.$$

Thus, $\lambda_w \leq \epsilon$. Since w was chosen arbitrarily, it follows that $\lambda^{(n)} \leq \epsilon$, and hence all rates $R < C_{\text{BSC}}$ are achievable.

(d) Clearly the overall capacity can be no greater than the smaller the subchannel capacities, i.e.,

$$C \leq \min(C_{\text{BSC}}(p_1), C_{\text{BSC}}(p_2)).$$

Now, choose codes C_1^* and C_2^* for the two subchannels both of which operate at rate $R < \min(C_{\text{BSC}}(p_1), C_{\text{BSC}}(p_2))$. Then for n sufficiently large, the codes can be designed to operate at an arbitrarily small maximal error rate. Using the same argument as above, we see that the error rate at the receiver can be made arbitrarily small, and so all rates $R < \min(C_{\text{BSC}}(p_1), C_{\text{BSC}}(p_2))$ are achievable, and hence the capacity is $C = \min(C_{\text{BSC}}(p_1), C_{\text{BSC}}(p_2))$.