Introduction to Information and Coding Theory November 15, 2017
(Prof. Permuter Haim, Mr. Oron Sabag and Mr. Iddo Gattengo)

Final Exam - Moed B
Total time for the exam: 3 hours!

Important: For **True / False** questions, copy the statement to your notebook and write clearly true or false. You should prove the statement if true, and provide counterexample otherwise.

1) **True or False (24 Points)**:
   a) **True/False**: For two random variables, X and Y, $H(f(X,Y)) \leq H(g(X)) + H(h(Y))$, where f, g, h are arbitrary functions. (6 pts)
   Solution: False.
   Let us assume that: $f(X,Y) = X, g(X) = 0, h(Y) = 0$.
   Hence: $H(f(X,Y)) = H(X) \geq 0 = H(0) + H(0) = H(g(X)) + H(h(Y))$.
   b) Consider a Gaussian channel where the input, X, has a power constraint P, the noise, $Z \sim N(0,1)$ and the output is $Y = X + Z$. The output Y is fed through a function $f_i(y) = y^i$ where i is an integer. The capacity of this channel is denoted by C_i. Complete $<, >, =$ between C_2 and C_4, prove your answer. (6 pts)
   Solution: $C_2 = C_4$.
   Notice that if we know $f_2(y)$ then we know $f_4(y)$, and vice versa. Therefore:

$$\max_{f(x):E(X^2)\leq P} I(X;Y^2) = \max_{f(x):E(X^2)\leq P} I(X;Y^4) \tag{1}$$

   c) Consider a clean channel with $|\mathcal{X}|$ inputs and outputs (Fig. 1). Two systems are defined as follows:
   System A: At each time, a random variable $Z \sim$ Unif$(1, \dots, |\mathcal{X}|)$ determines how many links can be used at the next channel use. This random variable is known to the encoder and the decoder. The capacity of this system is denoted by C_A.
   System B: In this system, there is a clean channel but with $|\mathcal{X}'| = \frac{1+\cdots+|\mathcal{X}|}{|\mathcal{X}|}$ inputs (the average amount of inputs) at all times. The capacity of this channel is denoted by C_B.
   True/False: The capacity of system B is larger then the capacity of system A, i.e. $C_B \geq C_A$.(12 pts)

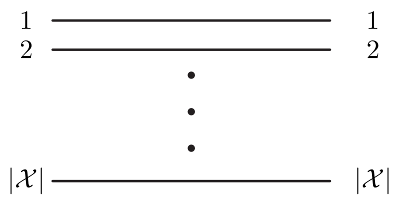


Fig. 1: Clean channel with $|\mathcal{X}|$ inputs.

Solution: True.
The capacity of a clean channel with $|\mathcal{X}|$ inputs is $\log|\mathcal{X}|$.
Applying it to System B we get:

$$\begin{aligned} C_B &= \log|\mathcal{X}'| \\ &= \log(\frac{1+\cdots+|\mathcal{X}|}{|\mathcal{X}|}) \\ &= \log[E[Z]] \end{aligned}$$

On the other hand: $C_A = \max_{p(x)} I(X;Y|Z)$, while:

$$I(X;Y|Z) = \sum_{i=1}^{|\mathcal{X}|} P(Z=i)I(X;Y|Z=i)$$

therefore,

$$\begin{aligned} C_A &= \frac{1}{|\mathcal{X}|}\sum_{i=1}^{|\mathcal{X}|}\log(i) \\ &= E[\log(Z)] \\ &\overset{(a)}{\leq} \log[E[Z]] \\ &= C_B \end{aligned}$$

Where (a) follows from Jensen's inequality.

2) **Constrained Markov chain (24 Points):**
A random process, $X_1, X_2, \dots$ is a Markov chain if it has the Markov property $X_i - X_{i-1} - X^{i-2}$ for all $i \geq 3$. In this question, the Markov chain $X_1, X_2, \dots$ takes values from a binary alphabet, $\mathcal{X} = \{0, 1\}$, and does not contain two consecutive ones (that is, $'11'$ is not valid). The conditional probability, $P_{X_i|X_{i-1}}$, of the Markov chain is given by

$$T = \begin{pmatrix} 1-p & p \\ 1 & 0 \end{pmatrix},$$

for all $i \geq 2$, where $p \in [0,1]$. The matrix rows correspond to X_{i-1} and the matrix columns correspond to X_i, for example, $P(X_i = 1|X_{i-1} = 0) = p$. The distribution of X_1 is to be defined later.

a) Explain why the Markov chain does not contain consecutive ones.
Solution:
From the conditional probability matrix: $P(X_i = 1|X_{i-1} = 1) = 0$.

b) The stationary distribution of a Markov chain is defined as a probability vector that solves $vT = v$. Find the stationary distribution of this Markov chain as a function of p.
Solution:
Assuming $v = \begin{bmatrix} v_1 & v_2 \end{bmatrix}$ we get: $v_1 + v_2 = 1$, and $v_1(1+p) = 1$. Hence: $v = \begin{bmatrix} \frac{1}{1+p} & \frac{p}{1+p} \end{bmatrix}$.

\- **From now on, assume that X_1 is distributed according to v that you found in (b)**.

c) Compute $P(X_2 = 0)$, $P(X_3 = 0)$ and $P(X_7 = 0)$ as a function of p.
Solution:
In this case, the probability vector is distributed the same for all $i \geq 1$, for:
$v_1 T = v_2, v_1 = v_2$ and $v_2 T = v_3, v_2 = v_3$ etc.
Hence, $P(X_2 = 0) = P(X_3 = 0) = P(X_7 = 0) = v_1 = \frac{1}{1+p}$.

d) (**True/False**) The entropy rate is defined as $H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} H(X^n)$. Is it true that $H(\mathcal{X}) = H(X_2|X_1)$?
Solution: True

$$\begin{aligned}
H(\mathcal{X}) &\overset{(a)}{=} \lim_{n\to\infty} \frac{1}{n} H(X^n) \\
&\overset{(b)}{=} \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(X_i|X^{i-1}) \\
&\overset{(c)}{=} \lim_{n\to\infty} \frac{1}{n} H(X_1) + \frac{1}{n} \sum_{i=2}^{n} H(X_i|X_{i-1}) \\
&= \lim_{n\to\infty} \frac{1}{n} H(X_1) + \frac{1}{n} \sum_{i=2}^{n} H(X_2|X_1) \\
&= \lim_{n\to\infty} \frac{1}{n} H(X_1) + \frac{n-1}{n} H(X_2|X_1) \\
&= H(X_2|X_1)
\end{aligned}$$

Where (a)-(c) follow from: chain rule, Markov property and stationary distibution, respectively.

e) Compute the entropy rate of the Markov chain as a function of p. (The answer should not contain a limit)
Solution:
From previous section we know that:

$$\begin{aligned}
H(\mathcal{X}) &= H(X_2|X_1) \\
&= P(X_1 = 0)H(X_2|X_1 = 0) + P(X_1 = 1)H(X_2|X_1 = 1) \\
&= v_1 H(p) + v_2 \cdot 0 \\
&= \frac{H(p)}{1+p}
\end{aligned}$$

f) In order to maximize the entropy rate, you can now optimize the parameter p. Does the optimal parameter satisfy $p = 0.5$, $p < 0.5$ or $p > 0.5$? (You don't have to solve the maximization problem but you should prove your answer.)

* Roughly speaking, the amount of sequences of length n and without $'11'$ is $2^{nH(\mathcal{X})}$. In magnetic storage, such as standard hard disk, it is useful to encode data into constrained sequences (in order to do decrease errors appearances) so the larger the entropy so the better it is.
Solution:
$H(p)$ is symmetric around $p = 0.5$. $1 + p$ is monotonically increasing. Hence we obviously prefer $p \leq 0.5$. Now substituting $p = 0.5$ we have: $H(\mathcal{X}) = \frac{2}{3}$. Let us check another input a bit smaller than 0.5, e.g. $p = 0.4$, we have $H(\mathcal{X}) = 0.694$. Now we may infer that $p < 0.5$.

3) **Polarization and the idea of polar codes (28 Points)**: The question is about polarization effect in memoryless channels that can lead to simple coding schemes that achieve the capacity which are called polar codes.

a) Consider the channel in Fig. 2 where two parallel binary erasure channels can be used at once (the input is $X = (X_1, X_2)$). The inputs alphabets are binary, so that Y_1 and Y_2 are the outputs of a $\mathrm{BEC}(p)$ with inputs X_1 and X_2, respectively.

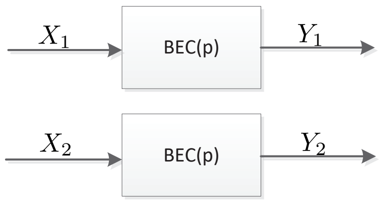


Fig. 2: Two parallel binary erasure channels

Compute the capacity of this channel, namely,

$$\max_{p(x_1,x_2)} I(X_1, X_2; Y_1, Y_2). \tag{2}$$

What is the input distribution $p(x_1, x_2)$ that achieves the capacity?

Solution:

$$\begin{aligned}
I(X_1, X_2; Y_1, Y_2) &= H(Y_1, Y_2) - H(Y_1, Y_2|X_1, X_2) \\
&= H(Y_1) + H(Y_2|Y_1) - (H(Y_1|X_1, X_2) + H(Y_2|X_1, X_2, Y_1)) \\
&= H(Y_1) + H(Y_2|Y_1) - (H(Y_1|X_1) + H(Y_2|X_2)) \\
&= H(Y_1) - H(Y_1|X_1) + H(Y_2|Y_1) - H(Y_2|X_2) \\
&\leq H(Y_1) - H(Y_1|X_1) + H(Y_2) - H(Y_2|X_2) \\
&= I(X_1, Y_1) + I(X_2, Y_2)
\end{aligned}$$

While equality holds if Y_1 and Y_2 are independent, that holds if X_1 and X_2 are independent.
As we saw in class, the capacity of a single $BEC(p)$ with input X and output Y is given by $C = (1-p)\sup_{Px} H(X) = 1 - p$, with $X \sim Bern(0.5)$ achieving it.
Hence the capacity of the described channel is given by:

$$\begin{aligned}
C &= \max_{p(x_1,x_2)} I(X_1, Y_1) + I(X_2, Y_2) \\
&= \max_{p(x_1)} I(X_1, Y_1) + \max_{p(x_2)} I(X_2, Y_2) \\
&= 2(1-p)
\end{aligned}$$

which is achieved by $X_1 \sim Bern(0.5), X_2 \sim Bern(0.5)$ and independent X_1 and X_2.

b) Consider the system in Fig. 3, where addition is modulo 2:

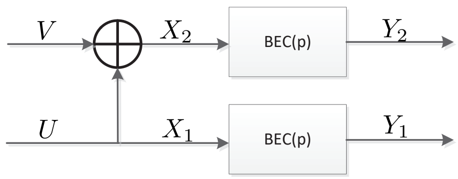


Fig. 3: Two parallel binary erasure channels with modified inputs

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} U \\ V \end{pmatrix}$$

Compute the capacity of the new channel, i.e. $\max_{p(u,v)} I(U, V; Y_1, Y_2)$.
What is the $p(u, v)$ that achieves the capacity?

Solution: U and V are functions of X_1,X_2. As a result,

$$I(U, V; Y_1, Y_2) \leq I(X_1, X_2; Y_1, Y_2). \tag{3}$$

With equality if X_1 and X_2 are functions of U,V (We know that $U = X_1$, and $X_2 = U \oplus V$). Now, the maximum mutual information of the new channel equals the capacity of the previous channel if we guarantee again that X_1,X_2 are independent and both are $\sim Bern(0.5)$, as it was shown previously. $V \sim Bern(0.5)$ establishes independence between X_1,X_2, and that $X_2 \sim Bern(0.5)$. Of course, $U \sim Bern(0.5)$ establishes $X_1 \sim Bern(0.5)$ since $U = X_1$.
Hence the new capacity is $C = 2(1-p)$, again.

Next, the channel is decomposed into two parallel channels as appears in Fig. 4. The input of Channel 1 is U and its output is (Y_1, Y_2, V). The input of Channel 2 is V and its output is (Y_1, Y_2).

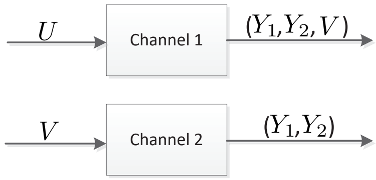


Fig. 4: Two new channels

c) Compute the expressions $I(U;Y_1,Y_2,V)$ and $I(V;Y_1,Y_2)$ with respect to the $p(u,v)$ that achieves the maximum in (b). What is the sum of the expressions you computed?
Solution:
Channel 1:

$$\begin{aligned}
I(U;Y_1,Y_2,V) &= I(U;Y_1) + I(U;Y_2|Y_1) + I(U;V|Y_1,Y_2) \\
&= (1-p)H(U) + H(U|Y_1) - H(U|Y_1,Y_2) + H(U|Y_1,Y_2) - H(U|Y1,Y2,V) \\
&= (1-p)H(U) + H(U|Y_1) - H(U|Y_1,Y_2,V)
\end{aligned}$$

While:

$$\begin{aligned}
H(U/Y_1) &= P(y_1 = '?')H(U|y_1 = '?') = pH(U) \\
H(U|Y_1,Y_2,V) &= p(y_1 = '?', y_2 = '?')H(U|V) = p^2H(U)
\end{aligned}$$

Hence:

$$\begin{aligned}
I(U;Y_1,Y_2,V) &= (1-p)H(U) + pH(U) - p^2H(U) \\
&= H(U) - p^2H(U) \\
&= H(U)(1-p^2)
\end{aligned}$$

Substituting $p(u) = 0.5$ we have:

$$I(U;Y_1,Y_2,V) = \quad 1-p^2$$

Channel 2:

$$\begin{aligned}
I(V;Y_1,Y_2) &= I(V;Y_1) + I(V;Y_2|Y_1) \\
&= H(V) - H(V|Y_1) + H(V|Y_1) - H(V|Y_1,Y_2) \\
&= H(V) - H(V|Y_1,Y_2) \\
&= H(V) - [p(y_1 \neq '?', y2 \neq '?')H(V|y_1 \neq '?', y2 \neq '?') \\
&+ p(y_1 \neq '?', y2 = '?')H(V|y_1 \neq '?', y2 = '?') \\
&+ p(y_1 = '?', y2 \neq '?')H(V|y_1 = '?', y2 \neq '?') \\
&+ p(y_1 = '?', y2 \neq '?')H(V|y_1 = '?', y2 \neq '?')] \\
&= H(V) - [0 + p(1-p)H(V) + p(1-p)\cdot min\{H(V),H(U)\} + p^2H(V)] \\
&= H(V) - p(1-p)[H(V) + min\{H(V),H(U)\}] - p^2H(V)
\end{aligned}$$

Substituting $p(u) = 0.5$ and $p(v) = 0.5$ we get:

$$\begin{aligned}
I(V;Y_1,Y_2) &= 1 - 2p(1-p) - p^2 \\
&= (1-p)^2
\end{aligned}$$

Now let us sum both:

$$\begin{aligned}
I(U;Y_1,Y_2,V) + I(V;Y_1,Y_2) &= 1 - p^2 + (1-p)^2 \\
&= 2(1-p)
\end{aligned}$$

d) Compare the mutual information of Channels 1 and 2 with the capacity of a binary erasure channel (that is, write $<$, $>$ or $=$ with simple proof).

*For large n, repeating this decomposition n times, ends up in nc clean channels and in $n(1-c)$ totally noisy channels. This is the main idea of polar codes, which achieves capacity.

Solution:

As mentioned:

$$C(BEC(p)) = (1-p) \cdot \sup_{P_x} H(X)$$
$$= 1 - p$$

Channel 1:

$$I(U; Y_1, Y_2, V) = 1 - p^2$$
$$= (1-p)(1+p)$$

which is greater than $C(BEC(p))$ because $(1+p) > 1$.

Channel 2:

$$I(V; Y_1, Y_2) = (1-p)^2$$

where $(1-p) < 1$, as a result the mutual information is less than $C(BEC(p))$.

4) **Logistic Regression (24 Points)**:
Recall the sigmoid function, $\sigma(z) = \frac{1}{1+e^{-z}}$. The logistic regression classifier, that we've learned in class, is a binary classifier. The estimated probability $\hat{p}(x^{(i)}; \theta)$ is defined as

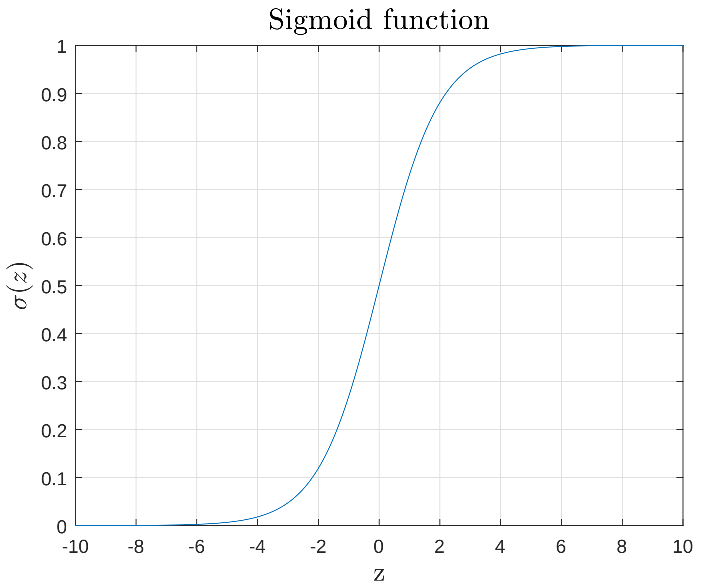
$$\hat{p}(y^{(i)} = 1 | x^{(i)}; \theta, b) = h_{\theta,b}(x^{(i)}) = f(\theta^\top x^{(i)} + b) \tag{-13}$$

where $f(z)$ is usually the sigmoid function $\sigma(z) = \frac{1}{1+e^{-z}}$.

a) Assume that $\{(x^{(i)}, y^{(i)})\}_{i=1}^m$ is a set of i.i.d samples Write the estimated probability of the entire set, i.e., write $\hat{p}\left(y^{(1)}, \dots, y^{(m)} | x^{(1)}, \dots, x^{(m)}\right)$ in terms of $\{(x^{(i)}, y^{(i)})\}_{i=1}^m$ and $h_{\theta,b}(\cdot)$.
Solution:

$$\hat{p}\left(y^{(1)}, \dots, y^{(m)} | x^{(1)}, \dots, x^{(m)}\right) = \prod_{i=1}^m h_{\theta,b}(x^{(i)})^{y^{(i)}} \left(1 - h_{\theta,b}(x^{(i)})\right)^{1-y^{(i)}}$$

b) Is the *sigmoid* function $\sigma(z)$ convex, concave or none? Prove your claim.
Solution: The sigmoid function is neither concave nor convex.



By drawing the sigmoid, it can be deduced that the function is not convex and not concave. Formally, a function is convex if $\lambda\sigma(z_1) + \bar{\lambda}\sigma(z_2) \geq \sigma(\lambda z_1 + \bar{\lambda} z_2)$. Set $z_1 = 0, z_2 = 10, \lambda = 0.8$ and the inequality fails. Similarly, it is not concave. Use $z_2 = -10$.

c) Assume that the sigmoid function is replaced with the following piecewise linear function

$$f(z) = \begin{cases} 0 & \text{if } z < -0.5 \\ 0.5 + x & \text{if } -0.5 \leq z \leq 0.5 \\ 1 & \text{if } z > 0.5 \end{cases} \tag{-12}$$

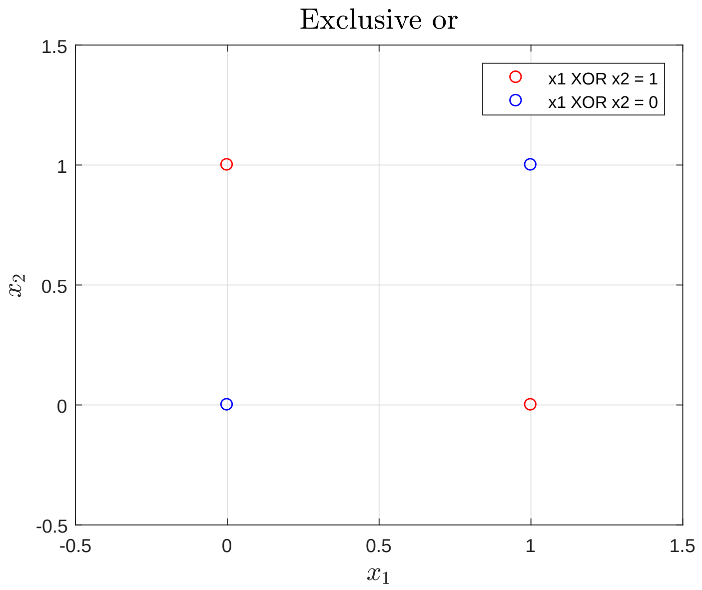
Let $x = (x_1, x_2)$ be a binary vector, namely, $x_1 \in \{0, 1\}, x_2 \in \{0, 1\}$. Can you find θ_1, θ_2 and b such that $f(\theta^\top x + b)$ is the logical or between x_1 and x_2? If yes, do it. If no, prove it doesn't exist. **Hint:** recall that $\theta^\top x = \theta_1 x_1 + \theta_2 x_2$.

Solution: Set $\theta_1 = \theta_2 = 1$ and $b = -0.5$. Then we have

$$f(x_1 + x_2 + -0.5) = \begin{cases} 0 & \text{if } (x_1, x_2) = (0, 0) \\ 1 & \text{otherwise} \end{cases}$$

d) Can you find θ, b such that $f(\theta^\top x + b)$ is the logical exclusive or (XOR) between x_1, x_2? If yes, do it. If no, prove it doesn't exist.

Solution: Such parameters doesn't exist.



We've learned that the logistic regression is a linear classifier - it defines a separating hyperplane with θ and b, and the classification is according to the sign of the inner product. Our function resembles sigmoid in that sense. There are no θ and b that will perform exclusive or because there is no such hyperplane in this case.

Good Luck!