Introduction to Information Theory and Machine Learning
(Prof. Permuter Haim, Mr. Bashar Huleihel and Ms. Yara Huleihel)

July 25th, 2023.

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

Please copy the following sentence and sign it: " I am respecting the rules of the exam: Signature:_________ "

Important: For **True / False** questions, copy the statement to your notebook and write clearly true or false. You should prove the statement if true, or disprove it, e.g. by providing a counter-example, otherwise.

1) **Transfer Entropy (36 Points):** Define the Transfer Entropy

$$\mathsf{TE}^{(k)}_{\mathcal{X}\to\mathcal{Y}}(t) = I\left(Y_t; X^{(k)}_{t-1} | Y^{(k)}_{t-1}\right), \tag{1}$$

where $X_t^{(k)} := (X_t, X_{t-1}, ..., X_{t-k+1})$ is a notation for length-k history of a variable X up to time t.
Let $\{X_t\}$ and $\{Y_t\}$ be stationary and first-order Markov processes taking values from the binary alphabet:

- Process $\{X_t\}$ has a deterministic transitions from 0 to 1 or 1 to 0 each time step, i.e.

$$P(X_t|Y^{t-1}, X^{t-1}) = P(X_t|X_{t-1}), \quad P(X_t = x|X_{t-1} = x \oplus 1) = 1, \tag{2}$$

where $P(X_0) \sim \text{Ber}\left(\frac{1}{2}\right)$.

- Process $\{Y_t\}$ is a noisy observation of the last time step of $\{X_t\}$. Assume $\alpha \neq \frac{1}{2}$ and $0 < \alpha < 1$,

$$P(Y_t|Y^{t-1}, X^{t-1}) = P(Y_t|X_{t-1}), \quad P(Y_t = y|X_{t-1} = x) = \begin{cases} 1-\alpha & \text{if } y = x \\ \alpha & \text{if } y \neq x \end{cases}. \tag{3}$$

Reminder: A stochastic process $\{X_t\}$ is said to be **stationary** if for every t_1, t_2 and h, the joint probability distribution function $P(X_{t_1}, X_{t_1+1}, ..., X_{t_1+h})$ is equal to $P(X_{t_2}, X_{t_2+1}, ..., X_{t_2+h})$, i.e., the joint probability distribution is invariant under time shifts.

a) **(8 points) True / False** The described joint process $\{X_t, Y_t\}$ is stationary. Explain your answer.
Solution: True.
Since $P(X_0) \sim \text{Ber}\left(\frac{1}{2}\right)$, we infer that $P(X_t) \sim \text{Ber}\left(\frac{1}{2}\right)$ for every t - thus X is markov that does not depend on t, and therefore the process X is stationary. Accordingly, we obtain that

$$P(X_t = x, Y_t = y) = \begin{cases} \frac{1}{2}(1-\alpha) & \text{if } y \neq x \\ \frac{1}{2}\alpha & \text{if } y = x \end{cases}. \tag{4}$$

It now follows that $P(X_{t_1}, Y_{t_1}) = P(Y_{t_1}|X_{t_1})P(X_{t_1})$. Specifically, $P(X_{t_1})$ is stationary and (3) reveals that $P(Y_{t_1}|X_{t_1})$ is not dependent on t.
Then we get that $P(X_{t_1}, Y_{t_1}, ..., X_{t_1+h}, Y_{t_1+h}) = P(X_{t_1}, ..., X_{t_1+h})P(Y_{t_1}, ..., Y_{t_1+h}|X_{t_1}, ..., X_{t_1+h})$ is stationary since $P(X_{t_1}, ..., X_{t_1+h})$ is stationary and $P(Y_{t_1}, ..., Y_{t_1+h}|X_{t_1}, ..., X_{t_1+h})$ is memoryless and does not depend on t from the definition in (4).

b) **(6 points) True / False** $P(Y_t = y, X_{t-1} = x) \neq P(X_t = x, Y_{t-1} = y)$.
Solution: False.
From the definition of the process we infer the followings,

$$\begin{aligned} P(Y_t = y, X_{t-1} = x) &= P(Y_{t-1} = y, X_{t-2} = x) && \text{(5a)} \\ &= P(Y_{t-1} = y, X_t = x), && \text{(5b)} \end{aligned}$$

where (5a) is from stationarity, and (5b) is from the definition of process $\{X_t\}$, since X_t is equal to X_{t-2}.

c) **(6 points)** Calculate the **Mutual Information** between Y_t and X_{t-1}, i.e. $I(Y_t; X_{t-1})$.
Hint: Consider to use the fact that $Y_t = X_{t-1} \oplus Z_{t-1}$, where $\{Z_t\}$ are i.i.d. Ber(α).
Solution: Consider a process $\{Z_t\}$ distributed i.i.d. Ber(α). Then,

$$\begin{aligned} I(Y_t; X_{t-1}) &= H(Y_t) - H(Y_t|X_{t-1}) && \text{(6a)} \\ &= H(Y_t) - H(X_{t-1} \oplus Z_{t-1}|X_{t-1}) && \text{(6b)} \\ &= H_b\left(\frac{1}{2}\right) - H(Z_{t-1}|X_{t-1}) && \text{(6c)} \\ &= 1 - H(Z_{t-1}) && \text{(6d)} \\ &= 1 - H_b(\alpha) && \text{(6e)} \\ &> 0, && \text{(6f)} \end{aligned}$$

where the last step is true because $\alpha \neq \frac{1}{2}$.

d) **(6 points) True / False** $I(Y_t; X_{t-1}) = I(X_t; Y_{t-1})$.
Solution: True.
We showed that $P(Y_t = y, X_{t-1} = x) = P(X_t = x, Y_{t-1} = y)$, and from the stationarity of the process $P(Y_t) = P(Y_{t-1}), P(X_t) = P(X_{t-1})$, and then we get,

$$I(Y_t; X_{t-1}) = \sum P(Y_t, X_{t-1}) \log \frac{P(Y_t, X_{t-1})}{P(Y_t)P(X_{t-1})} \tag{7a}$$

$$= \sum P(Y_{t-1}, X_t) \log \frac{P(Y_{t-1}, X_t)}{P(Y_{t-1})P(X_{t-1})} = I(X_t; Y_{t-1}). \tag{7b}$$

e) **(6 points)** Show that the Transfer Entropy for $X \to Y$ with lag $k = 1$ is non-zero, i.e., $\mathsf{TE}^{(1)}_{\mathcal{X}\to\mathcal{Y}}(t) = I\left(Y_t; X_{t-1}\middle|Y_{t-1}\right) > 0$.
Hint: Utilize the relation $Y_t = X_{t-1} \oplus Z_{t-1}$, and the fact that if $Z_1 \sim \text{Ber}(\alpha)$ and $Z_2 \sim \text{Ber}(\beta)$, then $Z_1 \oplus Z_2 \sim \text{Ber}(\alpha - 2\alpha\beta + \beta)$.
Solution: Consider a process $\{\bar{Z}_t\}$ distributed i.i.d. Ber$(1-\alpha)$.

$$\begin{aligned}
\mathsf{TE}^{(1)}_{\mathcal{X}\to\mathcal{Y}}(t) &= I\left(Y_t; X_{t-1}\middle|Y_{t-1}\right) && \text{(8a)}\\
&= H(Y_t|Y_{t-1}) - H(Y_t|Y_{t-1}, X_{t-1}) && \text{(8b)}\\
&= H(X_{t-1} \oplus Z_{t-1}|X_{t-2} \oplus Z_{t-2}) - H(Y_t|X_{t-1}) && \text{(8c)}\\
&= H(X_{t-1} \oplus 1 \oplus Z_{t-1} \oplus 1|X_{t-2} \oplus Z_{t-2}) - H_b(\alpha) && \text{(8d)}\\
&= H(X_{t-2} \oplus \bar{Z}_{t-2}|X_{t-2} \oplus Z_{t-2}) - H_b(\alpha) && \text{(8e)}\\
&= H(X_{t-2} \oplus \bar{Z}_{t-2} \oplus X_{t-2} \oplus Z_{t-2}|X_{t-2} \oplus Z_{t-2}) - H_b(\alpha) && \text{(8f)}\\
&= H(\bar{Z}_{t-2} \oplus Z_{t-2}|X_{t-2} \oplus Z_{t-2}) - H(\alpha) && \text{(8g)}\\
&= H(\bar{Z}_{t-2} \oplus Z_{t-2}) - H(\alpha) && \text{(8h)}\\
&= H_b(2\alpha^2 - 2\alpha + 1) - H_b(\alpha) && \text{(8i)}\\
&> 0. && \text{(8j)}
\end{aligned}$$

Note that the transition from (8g) to the next equation is true due to the fact that X_{t-2} is distributed Ber$\left(\frac{1}{2}\right)$ thus Z_{t-2} independent of $X_{t-2} \oplus Z_{t-2}$, thus $H(\bar{Z}_{t-2} \oplus Z_{t-2}|X_{t-2} \oplus Z_{t-2}) = H(\bar{Z}_{t-2} \oplus Z_{t-2})$.
It's easy to prove that for the given $\alpha \in (0,1)\diagdown\{\frac{1}{2}\}$ the last equation holds - without loss of generality, assume $\alpha \in \left(0, \frac{1}{2}\right)$, due to the fact that $H_b(\alpha) = H_b(1-\alpha)$. Then we need to prove that $2\alpha^2 - 2\alpha + 1 > \alpha$, and for any $\alpha \in \left(0, \frac{1}{2}\right)$ this is true.

f) **(4 points)** Calculate the Transfer Entropy for $Y \to X$ with lag $k = 1$, i.e., $\mathsf{TE}^{(1)}_{\mathcal{Y}\to\mathcal{X}} = I\left(X_t; Y_{t-1}\middle|X_{t-1}\right)$.
Solution:

$$\begin{aligned}
\mathsf{TE}^{(1)}_{\mathcal{Y}\to\mathcal{X}}(t) &= I\left(X_t; Y_{t-1}\middle|X_{t-1}\right) && \text{(9a)}\\
&= H(X_t|X_{t-1}) - H(X_t|X_{t-1}, Y_{t-1}) && \text{(9b)}\\
&= 0. && \text{(9c)}
\end{aligned}$$

(9c) is correct since X transition autonomously and deterministically, thus $H(X_t|X_{t-1}) = 0$, and $H(X_t|X_{t-1}, Y_{t-1}) = 0$
Solution Insight - The concept of Transfer Entropy reveals valuable insights when analyzing the relationship between two processes, X and Y. Unlike mutual information, Transfer Entropy is not symmetric and can provide us with specific information about the synergy between the past of Y and its current state. Additionally, it uncovers the directional flow of information between X and Y, particularly in the context of Y's past. This makes Transfer Entropy a powerful tool for understanding the intricate dependencies and information transfer dynamics between these processes.

2) **ML algorithms (36 Points):** Figure 1 shows the end-to-end communication system considered in this question. This system takes as input a bit sequence denoted by $\boldsymbol{b}$, which is then mapped onto symbols, $\mathsf{s} \in \mathcal{S}$. The sequence of symbols is fed into a symbol modulator that maps each symbol into a constellation point $x \in \mathbb{C}$. Both the modulator and demodulator are implemented with neural networks, hence they are learnable with trainable parameters $\boldsymbol{\theta}_M$ and $\boldsymbol{\theta}_D$, respectively. The demodulator maps each received sample $y \in \mathbb{C}$ to a probability vector $\tilde{p}_{\boldsymbol{\theta}_D}(s|y)$ over the set of symbols $\mathcal{S}$, as illustrated in Fig.1. Finally, the sent bits are reconstructed from $\tilde{p}_{\boldsymbol{\theta}_D}(s|y)$ by the symbols-to-bits mapper. We also denote by $p_{\boldsymbol{\theta}_M}(s|y)$ the distribution induced by the system up to the point of the output channel (without the demodulator), which depends on the modulator parameters $\boldsymbol{\theta}_M$. We would like to approximate the true posterior distribution $p_{\boldsymbol{\theta}_M}(s|y)$ with the mapping defined by the demodulator $\tilde{p}_{\boldsymbol{\theta}_D}(s|y)$.
Given that the demodulator performs a classification task, the categorical cross-entropy is used as a loss function for training:

$$\mathcal{L}^*(\boldsymbol{\theta}_M, \boldsymbol{\theta}_D) \triangleq \mathbb{E}_{S,Y}\{-\log\left(\tilde{p}_{\boldsymbol{\theta}_D}(S|Y)\right)\} \tag{10}$$

We assume that each symbol $\mathsf{s} \in \mathcal{S}$ is uniquely mapped to a constellation point $x \in \mathbb{C}$ (1:1 mapping), allowing us to replace s with x in expressions.

a) **(5 points)** In our system, $p_{\boldsymbol{\theta}_M}(x)$ represents the true distribution of x, and note it depends on the modulator parameters, $\boldsymbol{\theta}_M$.

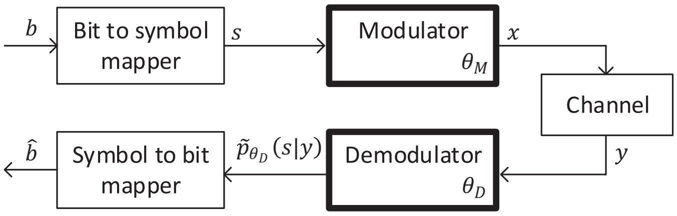

Fig. 1: Trainable end-to-end communication system. Trainable components are highlighted.

The entropy of X under the distribution parameterized by $\boldsymbol{\theta}_M$ is:

$$H_{\boldsymbol{\theta}_M}(X) = -\sum_x p_{\boldsymbol{\theta}_M}(x)\log\left(p_{\boldsymbol{\theta}_M}(x)\right) \tag{11}$$

True/False: $H_{\boldsymbol{\theta}_M}(X) = H(S)$? Explain why.
Solution:
True. 1:1 mapping. Each symbol is uniquely mapped to a constellation point $x \in \mathbb{C}$.

b) **(5 points)** Given a sequence of i.i.d. samples s_i and y_i over a long period of time, for $i = 1, 2, 3, ...$, how can you compute $\mathcal{L}^*(\boldsymbol{\theta}_M, \boldsymbol{\theta}_D)$?
Solution:
Given a sequence over a long period of time of i.i.d. samples s_i and y_i, for $i = 1, 2, 3, ...$, you can compute $\mathcal{L}^*(\boldsymbol{\theta}_M, \boldsymbol{\theta}_D)$ by taking the average of the negative log-likelihood of the predicted probabilities $\tilde{p}_{\boldsymbol{\theta}_D}(s_i|y_i)$ for the true symbols s_i given the received samples y_i, $-\frac{1}{L}\sum_{i=1}^{L}\log\left(\tilde{p}_{\boldsymbol{\theta}_D}(s_i|y_i)\right)$. As the number of samples goes to infinity, this average will converge to the expected value of the log loss function, $\mathbb{E}_{S,Y}\{-\log\left(\tilde{p}_{\boldsymbol{\theta}_D}(S|Y)\right)\}$. This is based on the law of large numbers, which states that as the number of i.i.d. samples increases, the sample average converges to the expected value.

c) **(5 points)** Recall the 1:1 mapping between s to x. Let:

$$\mathcal{L}(\boldsymbol{\theta}_M, \boldsymbol{\theta}_D) \triangleq \mathbb{E}_Y\{H\left[p_{\boldsymbol{\theta}_M}(x|y), \tilde{p}_{\boldsymbol{\theta}_D}(x|y)\right]\} \tag{12}$$

True/False? Is $\mathcal{L}(\boldsymbol{\theta}_M, \boldsymbol{\theta}_D)$ equals $\mathcal{L}^*(\boldsymbol{\theta}_M, \boldsymbol{\theta}_D)$, where $\mathcal{L}(\boldsymbol{\theta}_M, \boldsymbol{\theta}_D)$ is defined in (12), and $\mathcal{L}^*(\boldsymbol{\theta}_M, \boldsymbol{\theta}_D)$ in (10). If yes, prove it. Otherwise, correct the equation, and explain your reasoning.
Solution:

$$\begin{aligned}
\mathbb{E}_{S,Y}\{-\log\left(\tilde{p}_{\boldsymbol{\theta}_D}(S|Y)\right)\} &= \sum_y\sum_x p_{\boldsymbol{\theta}_M}(x,y)\log\left(\tilde{p}_{\boldsymbol{\theta}_D}(x|y)\right)\\
&= \sum_y\sum_x p_{\boldsymbol{\theta}_M}(y)p_{\boldsymbol{\theta}_M}(x|y)\log\left(\tilde{p}_{\boldsymbol{\theta}_D}(x|y)\right)\\
&= \mathbb{E}_Y\left\{\sum_x p_{\boldsymbol{\theta}_M}(x|y)\log\left(\tilde{p}_{\boldsymbol{\theta}_D}(x|y)\right)\right\}\\
&= \mathbb{E}_Y\left\{H\left[p_{\boldsymbol{\theta}_M}(x|y), \tilde{p}_{\boldsymbol{\theta}_D}(x|y)\right]\right\}.
\end{aligned}$$

d) **(5 points) True/False**? Is the loss function, $\mathcal{L}(\boldsymbol{\theta}_M, \boldsymbol{\theta}_D)$, as defined in (12) satisfies the following equality? If yes, explain why. If no, provide the correct expression.

$$\mathcal{L}(\boldsymbol{\theta}_M, \boldsymbol{\theta}_D) = -\sum_x\sum_y p_{\boldsymbol{\theta}_M}(x)p_{\boldsymbol{\theta}_M}(y|x)\log\left(\tilde{p}_{\boldsymbol{\theta}_D}(x|y)\right) \tag{13}$$

Solution:
True.

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}_M, \boldsymbol{\theta}_D) &= \mathbb{E}_Y\{H\left[p_{\boldsymbol{\theta}_M}(x|y), \tilde{p}_{\boldsymbol{\theta}_D}(x|y)\right]\}\\
&= -\sum_y p(y)\sum_x p_{\boldsymbol{\theta}_M}(x|y)\log\left(\tilde{p}_{\boldsymbol{\theta}_D}(x|y)\right)\\
&= -\sum_y\sum_x p_{\boldsymbol{\theta}_M}(x,y)\log\left(\tilde{p}_{\boldsymbol{\theta}_D}(x|y)\right)\\
&= -\sum_x\sum_y p_{\boldsymbol{\theta}_M}(x)p_{\boldsymbol{\theta}_M}(y|x)\log\left(\tilde{p}_{\boldsymbol{\theta}_D}(x|y)\right).
\end{aligned}$$

e) **(6 points) True/False**? Can the loss function be expressed as the following equation? If yes, prove it. Otherwise, correct the equation and explain your reasoning.

$$\mathcal{L}(\boldsymbol{\theta}_M, \boldsymbol{\theta}_D) = H(S) - I_{\boldsymbol{\theta}_M}(X;Y) + \mathbb{E}_Y\{D_{KL}\left(p_{\boldsymbol{\theta}_M}(x|y)\|\tilde{p}_{\boldsymbol{\theta}_D}(x|y)\right)\} \tag{14}$$

Solution:
True.

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}_M,\boldsymbol{\theta}_D) &= -\sum_x\sum_y p_{\boldsymbol{\theta}_M}(x)p_{\boldsymbol{\theta}_M}(y|x)\log\left(\tilde{p}_{\boldsymbol{\theta}_D}(x|y)\right)\\
&= -\sum_x\sum_y p_{\boldsymbol{\theta}_M}(x,y)\log\left(\frac{\tilde{p}_{\boldsymbol{\theta}_D}(x|y)p_{\boldsymbol{\theta}_M}(x)}{p_{\boldsymbol{\theta}_M}(x)}\right)\\
&= -\sum_x p_{\boldsymbol{\theta}_M}(x)\log\left(p_{\boldsymbol{\theta}_M}(x)\right) - \sum_x\sum_y p_{\boldsymbol{\theta}_M}(x,y)\log\left(\frac{\tilde{p}_{\boldsymbol{\theta}_D}(x|y)p_{\boldsymbol{\theta}_M}(y)}{p_{\boldsymbol{\theta}_M}(x)p_{\boldsymbol{\theta}_M}(y)}\right)\\
&= H_{\boldsymbol{\theta}_M}(X) - \sum_x\sum_y p_{\boldsymbol{\theta}_M}(x,y)\log\left(\frac{p_{\boldsymbol{\theta}_M}(x,y)}{p_{\boldsymbol{\theta}_M}(x)p_{\boldsymbol{\theta}_M}(y)}\right) - \sum_x\sum_y p_{\boldsymbol{\theta}_M}(x,y)\log\left(\frac{\tilde{p}_{\boldsymbol{\theta}_D}(x|y)}{p_{\boldsymbol{\theta}_M}(x|y)}\right)\\
&= H_{\boldsymbol{\theta}_M}(X) - I_{\boldsymbol{\theta}_M}(X;Y) - \sum_y p_{\boldsymbol{\theta}_M}(y)\sum_x p_{\boldsymbol{\theta}_M}(x|y)\log\left(\frac{\tilde{p}_{\boldsymbol{\theta}_D}(x|y)}{p_{\boldsymbol{\theta}_M}(x|y)}\right)\\
&= H(S) - I_{\boldsymbol{\theta}_M}(X;Y) + \mathbb{E}_Y\{D_{KL}\left(p_{\boldsymbol{\theta}_M}(x|y)\|\tilde{p}_{\boldsymbol{\theta}_D}(x|y)\right)\}.
\end{aligned}$$

f) **(5 points)** A student who saw equation (14) claims that:

$$\arg\min_{\boldsymbol{\theta}_M,\boldsymbol{\theta}_D}\hat{\mathcal{L}}(\boldsymbol{\theta}_M,\boldsymbol{\theta}_D) = \arg\min_{\boldsymbol{\theta}_M,\boldsymbol{\theta}_D}\mathcal{L}(\boldsymbol{\theta}_M,\boldsymbol{\theta}_D) \tag{15}$$

where $\hat{\mathcal{L}}(\boldsymbol{\theta}_M,\boldsymbol{\theta}_D)$ is defined as:

$$\hat{\mathcal{L}}(\boldsymbol{\theta}_M,\boldsymbol{\theta}_D) = \mathcal{L}(\boldsymbol{\theta}_M,\boldsymbol{\theta}_D) - H(S) \tag{16}$$

Is the student claim **True/False**? Please explain/justify your answer.
Solution:
True. The explanation is that $\mathcal{L}$ and $\hat{\mathcal{L}}$ differ by a constant, $H(S)$ doesn't depend on $\boldsymbol{\theta}_M$ or, $\boldsymbol{\theta}_D$.

g) **(5 points)** Please explain the contribution/meaning of minimizing the first loss component of $\hat{\mathcal{L}}(\boldsymbol{\theta}_M,\boldsymbol{\theta}_D)$, namely $-I_{\boldsymbol{\theta}_M}(X;Y)$, and the second loss component of $\hat{\mathcal{L}}(\boldsymbol{\theta}_M,\boldsymbol{\theta}_D)$, namely $\mathbb{E}_y\{D_{KL}\left(p_{\boldsymbol{\theta}_M}(x|y)\|\tilde{p}_{\boldsymbol{\theta}_D}(x|y)\right)\}$, to the overall communication system.
Solution:
Training the end-to-end system by minimizing $\hat{\mathcal{L}}$ corresponds to maximizing the mutual information of the channel inputs X and outputs Y, which is the capacity of the channel, while minimizing the KL divergence between the true posterior distribution $p_{\boldsymbol{\theta}_M}(x|y)$ and the one learned by the receiver $\tilde{p}_{\boldsymbol{\theta}_D}(x|y)$, trying to minimize the error.
Insight for this question: The goal of this task is to maximize the mutual information $I(X;Y)$ between the channel input X and the output Y, which is the capacity of the channel, by optimizing the constellation. Typically, finding the optimal $p(x)$ is challenging as it requires knowledge of the channel distribution $p(y|x)$. We aim to demonstrate that our objective can be achieved by minimizing the categorical cross-entropy.

3) **Polar compressor (32 Points):** For a positive integer N, let $n = 2^N$ and consider the invertible matrix $P_n \in \mathbb{F}_2^{n\times n}$ defined by:

$$P_n = \begin{bmatrix}1 & 0\\ 1 & 1\end{bmatrix}^{\otimes N}.$$

Further, consider $Z^n = (Z_1,\dots,Z_n) \sim Bern(p)^n$ where $p \in (0,0.5)$, and let $W^n = Z^n \cdot P_n$.

a) **(4 points)** For both channel coding and source coding, we do polarization. Explain briefly the difference in polarization between channel coding and source coding.
Solution:
The difference between the two problems lies in their objectives:

- Channel coding deals with transmitting information over a noisy communication channel. The main objective of channel coding is to introduce redundancy into the transmitted data so that errors introduced by the channel can be corrected or detected at the receiver. In the context of polar codes, besides the information bits we also consider frozen bits which indeed introduce redundancy,
- Source coding focuses on compressing the original data to reduce the number of bits required for transmission. The main objective of source coding is to efficiently represent the source data with fewer bits, thus achieving a lower compression ratio. In the context of polar codes, the encoder's operation involves multiplying the input data by the matrix P_n and then representing the message with fewer bits corresponding to those with the highest entropy, thus reducing data redundancy for compression.

b) **(4 points)** Define the rate for source coding and channel coding, and explain whether you want to maximize or minimize those rates.
Solution:
The rate in source coding refers to the average number of bits per symbol used to represent the source data. The lower the rate, the more efficient the compression, as it implies that fewer bits are needed to represent each symbol from the source.

The rate in channel coding refers to the ratio of useful information bits to the total number of transmitted bits, including the redundant bits. The lower the rate, the more redundant bits are used to protect the data, and consequently, the more resilient the communication system becomes against errors. However, a lower rate also means a lower throughput for the system, as more bits need to be transmitted, and therefore, our goal is to maximize the rate in channel coding, and minimize the rate in source coding.

c) **(6 points)** Assume $n = 4$ and consider the entropy terms $H(W_i|W^{i-1})$ for $i \in \{1, \ldots, 4\}$. Determine and explain which one is the highest, and calculate this specific entropy term explicitly in terms of p.

Solution:

Since $p \in (0, 0.5)$, we know that polarization occurs with the chosen polarization matrix P_4. Thus, the highest entropy term among the four is $H(W_1)$, which corresponds to applying the transform twice, each time taking the worst direction. To calculate $H(W_1)$, let us determine $P(W_1 = 1)$. Note that

$$W_1 = Z_1 \oplus Z_2 \oplus Z_3 \oplus Z_4,$$

where $\oplus$ denotes the XOR operation. Thus, $W_1 = 1$ if the sequence Z^4 consists of either exactly 3 zeros or exactly 3 ones. Accordingly, we can conclude that

$$P(W_1 = 1) = 4p^3(1-p) + 4(1-p)^3 p.$$

Thus,

$$H(W_1) = H_b(4p^3(1-p) + 4(1-p)^3 p).$$

d) **(6 points)** Define the set S_τ as follows:

$$S_\tau = \{i \in \{1, \ldots, 4\} \mid H(W_i|W^{i-1}) \geq \tau\}.$$

For $\hat{\tau} = -\mathbb{E}[\log_2(P_{W_1})]$, write explicitly the set $S_{\hat{\tau}}$. Explain your result.

Solution:

Note that $H(W_1) = -\mathbb{E}[\log_2(P_{W_1})]$, and therefore, $\hat{\tau}$ is exactly equal to $H(W_1)$. Further, we know that polarization occurs and $H(W_1)$ has the highest entropy among the 4 terms. Combining this with the fact that $\hat{\tau} = H(W_1)$, we can conclude that the set $S_{\hat{\tau}}$ includes only the index 1, i.e., $S_{\hat{\tau}} = 1$.

e) **(6 points)** Consider $z^4 = [1, 0, 1, 1]$ and the set $S_{\hat{\tau}}$ that you found in the previous item. What is the output of the encoder?

Solution:

Let us follow the operation of the encoder. The first step is multiplying z^4 by the invertible matrix P_4. This results in the following vector:

$$\begin{aligned} w^4 &= z^4 \cdot P_4 \\ &= \begin{bmatrix}1, 0, 1, 1\end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix} \\ &= \begin{bmatrix}1, 1, 0, 1\end{bmatrix}. \end{aligned}$$

Now since the set $S_{\hat{\tau}}$ consists only of the first index, then the output of the encoder is $w^{|S_{\hat{\tau}}|} = w_1 = 1$.

f) **(6 points)** This time let $n = 2$, and assume that $Z_1 \sim Bern(p_1)$ and $Z_2 \sim Bern(p_2)$ are sampled conditioned on $Z_1 + Z_2 = a$ (for $a \in \mathbb{F}_2$). Let $b(p_1, p_2, a)$ denote the probability of Z_2 being 1 conditioned on $Z_1 + Z_2 = a$. Find $b(p_1, p_2, a)$ for both $a = 0$ and $a = 1$.

Solution:

We calculate $b(p_1, p_2, 0)$ as follows:

$$\begin{aligned} b(p_1, p_2, 0) &= P(Z_2 = 1 | Z_1 + Z_2 = 0) \\ &= P(Z_1 = 1, Z_2 = 1 | Z_1 + Z_2 = 0) \\ &= \frac{P(Z_1 = 1, Z_2 = 1)}{P(Z_1 + Z_2 = 0)} \\ &= \frac{p_1 p_2}{p_1 p_2 + (1 - p_1)(1 - p_2)}. \end{aligned}$$

In the same manner we get:

$$\begin{aligned} b(p_1, p_2, 1) &= \frac{P(Z_1 = 0, Z_2 = 1)}{P(Z_1 + Z_2 = 1)} \\ &= \frac{(1 - p)p_2}{p_1(1 - p_2) + (1 - p_1)p_2}. \end{aligned}$$

Good Luck!