Introduction to Information Theory and Machine Learning August 14^{th}, 2025.
(Prof. Permuter Haim, Dr. Bashar Huleihel, Mr. Rom Hirsch, and Mr. Eitan Spivak)

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:_________ "

1) **Word-Guessing and Model Mismatch (35 Points):** You are playing a word-guessing game. There are 1024 possible words, all equally likely.
   - **(a) (5 points)** Before making any guesses, what is the entropy (uncertainty) of the random variable representing the word?
   - **(b) (10 points)** You guess the **first letter** of the word. The computer tells you whether you're correct. You are debating between two guesses:
     - **T**: 1 out of every 4 words starts with T.
     - **L**: 1 out of every 8 words starts with L.

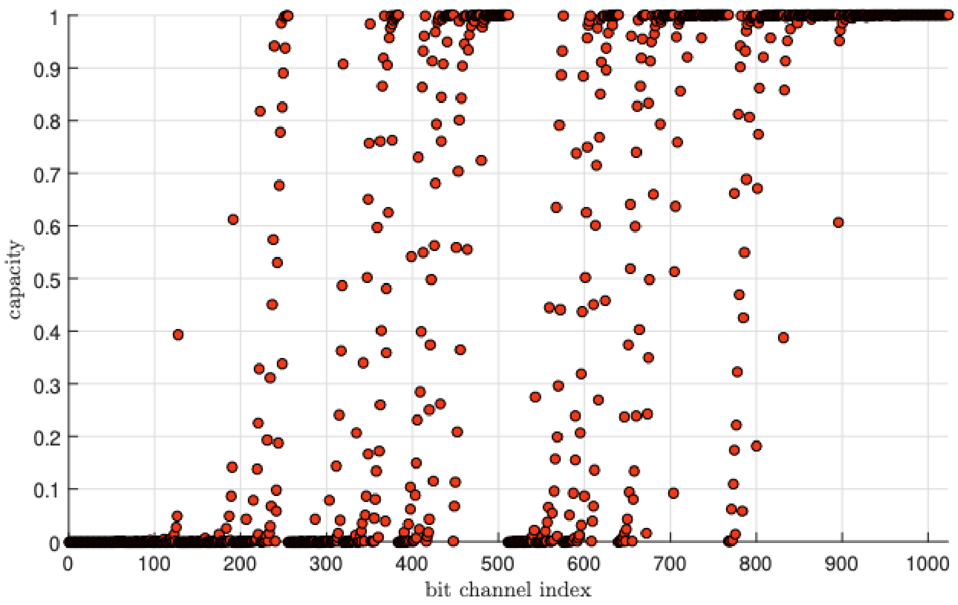
     Which letter gives you a better average reduction in entropy (uncertainty)? By how much?

     Remark: You may use the approximations
     $$\log_2(768) \approx 9.6, \quad \log_2(896) \approx 9.8.$$
   - **(c) (10 points)** Now consider the letter **R**, which starts **half of the words**. However, there is a twist: 10% of the time you guess R, the computer will not respond at all (but your guess will still be counted as used). Which is a better strategy: guessing R, or your better choice from part (b)? Justify.
   - **(d) (10 points)** You are given a file containing the letters $\{a, b, c, d\}$ whose empirical distribution is
     $$P = \left[\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right].$$
     You wish to compress this file using a code optimized for one of the following fixed probability models:
     $$Q_1 = \left[\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right], \quad Q_2 = \left[\frac{1}{2}, \frac{1}{8}, \frac{1}{8}, \frac{1}{4}\right].$$
     Which *information measure* should be used to determine which model (Q_1 or Q_2) yields better compression when the true source distribution is P? Compute its value for both Q_1 and Q_2, and state which model is better, with a brief explanation.

2) **Perfect Secrecy (35 Points):** Consider a communication scenario where Alice wants to send a message M, randomly drawn from a finite set $\mathcal{M}$, to Bob. To keep the message hidden from eavesdroppers, she encrypts it using a secret key $K \in \mathcal{K}$ that is known *only* to Alice and Bob and is independent of M. The encryption is performed using a deterministic function $C = f(K, M)$, producing encrypted message $C \in \mathcal{C}$. Bob decrypts the encrypted message using another deterministic function $M = g(K, C)$, and throughout the question we assume that such a decryption function g exists. The system is said to achieve *perfect secrecy* if $I(M; C) = 0$.
   - **(a) (5 points)** Briefly explain why a perfectly secure system is safe from an eavesdropper.
   - **(b) (7 points) True/False:** Under any system (whether secure or not), it holds that $H(M \mid C) \leq H(K \mid C)$. **Hint:** Use the assumption that there exits g such that $M = g(K, C)$.
   - **(c) (7 points)** Show that $I(M; C) \geq H(M) - H(K)$.
   - **(d) (6 points) True/False:** A student claims that, in order to achieve perfect secrecy, we must have $H(K) \geq H(M)$. If true, explain the meaning of this condition in no more than two lines. If false, provide a counterexample.
   - **(e) (10 points)** Now assume that M, K, and C are all n-bit binary strings, i.e., $M = K = C = \{0, 1\}^n$. Let M and K be independent and uniformly distributed over $\{0, 1\}^n$. Suggest encryption and decryption functions $f(K, M)$ and $g(K, C)$ that achieve perfect secrecy.

3) **Polar Code (35 Points):**

a) **(6 points) True/False:** Consider the graph below showing the capacities of the synthetic channels for a polar code of length $N = 1024$. A student claims that using a code rate of 0.8 will result in a block error probability very close to zero. Justify your answer in one or two sentences.



b) **(6 points)** We want to transmit the information bits $(1, 0)$ using a polar code of length $N = 4$ over a binary erasure channel (BEC). Select frozen bits to achieve the best decoding performance, and explain your choice. Then, compute the codeword (X_1, X_2, X_3, X_4).

c) **(7 points)** Assume the codeword from part (b) is sent over $\mathrm{BEC}(p)$ and the receiver observes $y = (?, ?, 1, 0)$. Perform successive cancellation (SC) decoding and show if the decoder succeeded in decoding the bits.
Remark: You may use the SC decoder functions: $g(r_1, r_2, b) = r_2 + (1 - 2b)r_1$ and $f(r_1, r_2) = \text{sign}(r_1)\text{sign}(r_2)\text{min}(|r_1|, |r_2|)$.

d) **(16 points)** We aim to develop an SC-based neural network model, where the goal is to learn the check-node function f_θ and the bit-node function g_θ. Suppose the analytic forms of f and g are unknown, but you are provided with a large dataset of input and log-likelihood ratio (LLR) pairs $D = \{(x_i, l_i)\}_{i=1}^M$, where $x_i \in \{0, 1\}$ are transmitted bits and $l_i \in \mathbb{R}$ are the corresponding LLR obtained after transmission through the channel, i.e., $l_i = \log\left(\frac{P(x_i=0|y_i)}{P(x_i=1|y_i)}\right)$, with y_i denoting the received channel output. The transmitted bits are independent and uniformly distributed, i.e., $P(x_i = 0) = P(x_i = 1) = 0.5$.

i) Describe how to generate a new training dataset for training f_θ and g_θ from $\{(x_i, l_i)\}_{i=1}^M$.
Hint: Recall f takes two LLRs and outputs one LLR; g takes two LLRs and a decoded bit, and outputs one LLR.

ii) Propose a method to learn f_θ and g_θ from the dataset in part (i), specifying the cost function, and provide a block diagram illustrating each model's inputs and outputs.

Good Luck!