Introduction to Information Theory and Machine Learning July 26th, 2026.
(Prof. Haim Permuter, Dr. Bashar Huleihel, Mr. Rom Hirsch, and Mr. Eitan Spivak)

Final Exam - Moed A
Total time for the exam: 3 hours, Total points: 125

1) **Cross-Entropy of Three Predictors (42 Points):** A student trains a model to guess a label $X \in \mathcal{X} = \{1, 2, \dots, M\}$ from a feature $Y \in \mathcal{Y}$. The training data and the test data come from one fixed joint distribution $P_{X,Y}$. A *predictor* is any conditional distribution $Q(x \mid y)$ on $\mathcal{X}$. We give each predictor a score on the test data: its (conditional) **cross-entropy loss**

$$\mathcal{L}(Q) \triangleq H\big(P_{X,Y},\, Q_{X|Y}\big) \triangleq -\sum_{x,y} P_{X,Y}(x,y) \log Q(x \mid y) = -\mathbb{E}[\log Q(X \mid Y)]\,,$$

with the convention $0 \log 0 = 0$. We look at three predictors:

$$Q_1(x \mid y) = 1/M, \qquad Q_2(x \mid y) = P_X(x), \qquad Q_3(x \mid y) = P_{X|Y}(x \mid y).$$

All logarithms are base 2. In parts (a)–(c), give the shortest answer you can, using only M, log, and the entropy symbols $H(\cdot)$ or $H(\cdot \mid \cdot)$.

a) **(4 points)** Find $\mathcal{L}(Q_1)$.
b) **(4 points)** Find $\mathcal{L}(Q_2)$.
c) **(4 points)** Find $\mathcal{L}(Q_3)$.
d) **(4 points) True or False:** for every joint $P_{X,Y}$,

$$\mathcal{L}(Q_1) \;\geq\; \mathcal{L}(Q_2) \;\geq\; \mathcal{L}(Q_3).$$

Explain in at most 2 lines.

e) **(4 points)** For each equality below, give the requested condition.
   (i) **(2 points)** $\mathcal{L}(Q_1) = \mathcal{L}(Q_2)$. Give the condition on P_X.
   (ii) **(2 points)** $\mathcal{L}(Q_3) = 0$. Give the condition on $P_{X,Y}$.

f) **(8 points)** Now the student can only see a simpler version of the feature: $V = g(Y)$, where $g : \mathcal{Y} \to \mathcal{V}$ is a fixed (possibly many-to-one) function. They build the predictor

$$Q_4(x \mid y) \triangleq P_{X|V}(x \mid g(y)).$$

What is the difference $\mathcal{L}(Q_4) - \mathcal{L}(Q_3)$? Write it in terms of mutual information, and interpret the result in the case $V = Y$ and in the case V is constant.

In parts (g) and (h), assume $P_{X,Y}(x,y) > 0$ and $Q(x \mid y) > 0$ for every $(x,y) \in \mathcal{X} \times \mathcal{Y}$.

g) **(4 points)** Let $(X,Y) \sim P_{X,Y}$, and for any predictor $Q(x \mid y)$ define

$$Z \triangleq \frac{Q(X \mid Y)}{P_{X|Y}(X \mid Y)}.$$

Find $\mathbb{E}[Z]$.

h) **(10 points) True or False:** for every predictor Q,

$$\mathcal{L}(Q) \;\geq\; \mathcal{L}(Q_3).$$

If the statement is true, provide a proof; if it is false, provide a counterexample.
(Hint: part (g) may be useful; e.g. apply Jensen's inequality to $-\log$.)

2) **Binary Channel with Two Noisy Observations (42 Points):**
Consider a memoryless channel with binary input $X \in \{0,1\}$ and output $Y = (Y_1, Y_2) \in \{0,1\}^2$. For each channel use,

$$Y_1 = X \oplus Z_1, \qquad Y_2 = X \oplus Z_2,$$

where $\oplus$ denotes addition modulo 2, and

$$Z_1, Z_2 \overset{\text{i.i.d.}}{\sim} \text{Bern}(p), \qquad 0 \leq p \leq \frac{1}{2},$$

independent of X.

a) **(8 points)** Characterize the channel as follows.
   (i) **(4 points)** Write the channel transition matrix

$$W(y_1, y_2 \mid x) = \Pr(Y_1 = y_1, Y_2 = y_2 \mid X = x).$$

   (ii) **(4 points)** Write the formula for the capacity of a general channel with input X and output Y. Then write the corresponding formula for the channel above, whose output is (Y_1, Y_2), and denote its capacity by $C_2(p)$.

b) **(6 points)** Find a capacity-achieving input distribution for this channel.
Hint: Examine the symmetry between the two rows of the transition matrix.

c) **(12 points)** Use the capacity-achieving input distribution found in the previous part.
   (i) **(4 points)** Compute $H(Y_1, Y_2 \mid X)$.
   (ii) **(5 points)** Find the probability distribution of (Y_1, Y_2) and compute

$$H(Y_1, Y_2).$$

(iii) **(3 points)** Compute $C_2(p)$ as a function of p.

d) **(8 points)** Assume that X is uniformly distributed over $\{0, 1\}$.

(i) **(4 points)** Determine the MAP estimate $\widehat{X}(y_1, y_2)$ for each of the four possible output pairs. If the two possible input values have equal posterior probabilities, choose between them uniformly at random.

(ii) **(4 points)** Compute the resulting probability of error

$$P_e = \Pr(\widehat{X} \neq X).$$

e) **(8 points)** Answer the following questions about processing the channel output.

(i) **(3 points)** Suppose that instead of observing the complete pair (Y_1, Y_2), the receiver observes only

$$T = Y_1 \oplus Y_2.$$

Compute the capacity of the induced channel from X to T.

(ii) **(3 points)** A receiver that observes only Y_1 sees a binary symmetric channel with crossover probability p, whose capacity is $C_1(p) = 1 - h_2(p)$. Compare $C_1(p)$ with the capacity $C_2(p)$ found in part (c).

(iii) **(2 points)** Evaluate $C_2(p)$ at $p = 0$ and at $p = \frac{1}{2}$, and briefly explain the operational meaning of each result.

3) **Convolutional Codes and Viterbi Decoding (41 Points):**

Consider a binary convolutional encoder with one input bit u_t at each time and two output bits $v_t = (v_t^{(1)}, v_t^{(2)})$. The generator sequences are

$$g^{(1)} = (1, 0, 1), \qquad g^{(2)} = (1, 1, 1).$$

Let $\mathbf{u} = (u_1, u_2, \ldots)$. For generator $g^{(i)}$, define $v_t^{(i)} = (\mathbf{u} * g^{(i)})_t = \sum_{\ell=0}^{m} g_\ell^{(i)} u_{t-\ell}$, where $*$ is convolution over $\mathbb{F}_2$. The encoder starts in the all-zero state.

a) **(5 points)** Draw the encoder diagram. Define the convolutional code parameters (n, k, m), its rate, and the number of states. Define the state s_t and the output vector v_t.

b) **(5 points)** Show the state transitions, either as a table or as a state diagram. Each branch should be labeled by

input/output.

c) **(5 points)** We want to transmit the information bits

$$(1, 1, 0)$$

using the convolutional code above. Assume the decoder knows that the final state is 00. Write the serialized transmit vector

$$V = (v_1^{(1)}, v_1^{(2)}, v_2^{(1)}, v_2^{(2)}, \ldots).$$

What is the length of V?

d) **(18 points)** For this part, the convolutional encoder above is used to transmit two information bits. The resulting coded bits are mapped to BPSK symbols by

$$0 \mapsto +1, \qquad 1 \mapsto -1.$$

The symbols are transmitted over an AWGN channel

$$Y_i = X_i + N_i, \qquad N_i \sim \mathcal{N}(0, \sigma^2),$$

with $\sigma^2 = 1$. The receiver observes

$$y = (1, 1,\ 1, -2,\ -2, 1,\ 1, 1).$$

The squared Euclidean metric is

$$d_E^2(x, y) = \sum_i (y_i - x_i)^2.$$

Assume the trellis path starts and ends in state 00. **Explain your answer in each part.**

(i) **(6 points)** Apply hard decisions to the received samples, and then use Viterbi decoding with Hamming-distance branch metrics. What are the estimated transmitted information bits? Explain your answer.

(ii) **(8 points)** Apply Viterbi decoding using the squared Euclidean metric. What are the estimated transmitted information bits? Explain your answer.

(iii) **(4 points)** Compare Viterbi decoding with the Hamming metric and with the squared Euclidean metric. Which metric is better for this received vector, and why can the two decoders give different estimated information bits? Explain your answer.

e) **(8 points)** Answer briefly:

(i) What is the difference between what the Viterbi algorithm outputs and what the BCJR algorithm computes?

(ii) What is the decoding algorithm used in a turbo-code decoder: BCJR, Viterbi as in part (d), or can both be used? Explain why in one or two lines.

Good Luck!