EE 276 - Information Theory
Final
March 19, 2024

1. There are a total of 5 questions. You have 3 hours to take the exam. Questions vary in difficulty and number of points. There are a total of 100 points.
2. Please write all answers in the designated area underneath the question. If you need more room for your answer, please indicate that under the question, and continue your response elsewhere.
3. Scratch paper will be provided and collected at the end of the exam, but will **not** be graded.
4. Answers should be justified, unless otherwise stated.
5. You are allowed to use non-electronic notes and material.
6. Calculators or any other electronic devices are not allowed.

Good luck!

Name:

SUID:

1. (*14 points*) Answer true or false to each of the following. You do not need to provide an explanation. A correct answer gets 2 points. An incorrect answer gets 0 points. Leaving the answer blank gets 1 point.

		T	F
(a)	Shannon and Huffman coding both assume knowledge of the source distribution.	X	
(b)	The discrete entropy of a random variable is invariant under one-to-one transformations of the random variable but the differential entropy is generally not.	X	
(c)	If the mutual information $I(X;Y) = 0$, then for any Z we have $I(X;Y\|Z) = 0$.		X
(d)	The noisy channel coding theorem guarantees that we can achieve a probability of error 0 for any block length n, as long as the rate is below the capacity of the channel.		X
(e)	Relative entropy between two distributions satisfies $D(p\|q) = D(q\|p)$.		X
(f)	The entropy rate of any stationary process $X_1, X_2, \ldots$ is never larger than that of the i.i.d. process $Y_1, Y_2, \ldots$, if X_1 and Y_1 have the same distribution.	X	
(g)	Recall $\mathbb{P}^n$ denotes the set of PMFs that are empirical distributions of sequences of length n with components in a given finite alphabet. If $p, q \in \mathbb{P}^n$, and $p \neq q$, then $T(p) \cap T(q) = \emptyset$, where $T(p)$ is the type class of p.	X	

2. (*17 points*) Consider a PMF p over an alphabet $\mathcal{X}$ of size 4 given by the following table

x	a	b	c	d
$p(x)$	1/2	1/4	1/8	1/8

(a) Suppose random variable X is drawn from p but not observed. To discover the value of X, you're allowed to ask a sequence of 'yes/no' questions that will be truthfully answered. What is the minimum expected number of such questions that you could ask in order to discover the selected symbol? What is the sequence of questions achieving the minimum? Explain your reasoning.

Solution: The series of questions "Is it A?", "Is it B?", etc. is optimal because each question reduces uncertainty by $\frac{1}{2}$, or 1 bit. The expected number of questions given by this scheme is:

$$\begin{aligned}\mathbb{E}\left[\#\text{ of questions}\right] &= 1 * \frac{1}{2} + 2 * \frac{1}{4} + 3 * \frac{2}{8} \\ &= \frac{1}{2} + \frac{1}{2} + \frac{6}{8} \\ &= \frac{7}{4} = 1.75\end{aligned}$$

(b) What is the entropy of X from the previous part?
Solution:

$$\begin{aligned}H &= -\sum_{i=1}^{8} p_i \log_2 p_i = \sum_{i=1}^{3} n * 2^{-n} + 3 * 2^{-3} \\ &= \frac{7}{4} = 1.75\end{aligned}$$

A code $C(\cdot)$ for the alphabet $\mathcal{X}$ is *non-singular* if for all $x_1, x_2 \in \mathcal{X}$,

$$x_1 \neq x_2 \quad \text{implies} \quad C(x_1) \neq C(x_2).$$

In words, it doesn't assign the same codeword to two different alphabet symbols.

(c) For the alphabet $\mathcal{X}$ above, provide an example of a non-singular code which is not uniquely decodable. Explain why your answer satisfies the requirements.
Solution: Consider the following example:

A	B	C	D
0	00	01	10

Each symbol in the alphabet (A-D) has a unique codeword. However, this code is not uniquely decodable because we have no way of decoding the bit string "00" to the character string AA or B.

(d) Construct a code which is uniquely decodable, but not a prefix code. Explain why it satisfies these requirements.

Soultion: The following is uniquely decodable because the codewords for the symbols A-D are the same as problem 3 on HW2. We can decode them using the same 0 counting argument we used on that homework.

A	B	C	D
00	10	11	110

(e) Construct a uniquely decodable prefix-free code which is optimal in the sense of expected code length for the source above. Explain why it satisfies these requirements.

Soultion: Using Huffman coding, we can generate

A	B	C	D
0	10	110	111

We know that Huffman codes are uniquely-decodable, prefix-free codes. You can see that no codeword is a prefix of any other. Huffman codes are optimal for dyadic distributions. We can see this by noticing:

$$E[\ell] = 1 * \frac{1}{2} + 2 * \frac{1}{4} + 3 * \frac{1}{8} + 3 * \frac{1}{8} = \frac{7}{4} = H(X)$$

(f) Relate your prefix code to the 'yes/no' questioning scheme you proposed in (a).
Solution The bits represent our history of yes/no (1/0) responses to our questions when the answer is a given symbol.

3. (*14 points*) Let x^n be a sequence of n elements each taking values in a finite alphabet $\mathcal{X}$. Let $\epsilon > 0$, and p, q be two pmfs on $\mathcal{X}$. We say the sequence x^n is "*divergence ϵ-typical for p relative to q*" if

$$\left| \frac{1}{n} \log \frac{p(x^n)}{q(x^n)} - D(p\|q) \right| \leq \epsilon.$$

We'll use $\mathcal{A}_\epsilon^{(n)}(p|q)$ to denote the set of all such sequences.

(a) Find

$$\lim_{n\to\infty} P(\mathcal{A}_\epsilon^{(n)}(p|q)),$$

where $P(\mathcal{A})$ denotes the probability of the event $\{X^n \in \mathcal{A}\}$ when X_i are drawn IID from p.

Solution: Note that the event in question is the same as

$$\left| \frac{1}{n} \sum_{i=1}^{n} \log \frac{p(X_i)}{q(X_i)} - \mathbb{E}\left[\log \frac{p(X)}{q(X)} \right] \right| \leq \epsilon \tag{1}$$

where the expectation is taken with respect to p. Since X_i are IID, the law of large number applies and gives that, for any ϵ, the probability of this event converges to 1.

(b) Show that

$$(1-\epsilon)2^{-n(D(p\|q)+\epsilon)} \leq Q(\mathcal{A}_\epsilon^{(n)}(p|q)) \leq 2^{-n(D(p\|q)-\epsilon)}$$

for n sufficiently large, where $Q(\mathcal{A})$ denotes the probability of the event $\{X^n \in \mathcal{A}\}$ when X_i are drawn IID from q.

Solution:

From the definition, by re-arranging we conclude that x^n is in the set of interest if and only if

$$D(p\|q) - \epsilon \leq \frac{1}{n}\log\frac{p(x^n)}{q(x^n)} \leq D(p\|q) + \epsilon$$

and hence

$$p(x^n)2^{-n(D(p\|q)+\epsilon)} \leq q(x^n) \leq p(x^n)2^{-n(D(p\|q)-\epsilon)}.$$

Summing over all $x^n \in \mathcal{A}_\epsilon^{(n)}(p|q)$ gives

$$P(\mathcal{A}_\epsilon^{(n)}(p|q))2^{-n(D(p\|q)+\epsilon)} \leq Q(\mathcal{A}_\epsilon^{(n)}(p|q)) \leq P(\mathcal{A}_\epsilon^{(n)}(p|q))2^{-n(D(p\|q)-\epsilon)}.$$

Now $P(\mathcal{A}_\epsilon^{(n)}(p|q)) \leq 1$ since it's a probability, meanwhile, from part (a), $P(\mathcal{A}_\epsilon^{(n)}(p|q)) \to 1$ as $n \to \infty$ and hence for any $\epsilon > 0$, there exists an n such that

$$P(\mathcal{A}_\epsilon^{(n)}(p|q)) \geq (1-\epsilon).$$

Combining these bounds on $P(\mathcal{A}_\epsilon^{(n)}(p|q))$ with the previous display completes the proof.

4. (*35 points*) For $0 \leq q \leq 1$, let

$$\phi(q) = \max H(W),$$

where the maximization is over all probability mass functions of the random variable W with alphabet $\mathcal{W} = \{0, 1, 2\}$ and satisfying $P(W \neq 0) \leq q$.

(a) Evaluate and qualitatively plot the function $\phi(q)$. Hint: for $0 \leq q \leq 2/3$ show that the maximum is achieved by the random variable W_q distributed as

$$W_q = \begin{cases} 0 & w.p. \ \ 1-q \\ 1 & w.p. \ \ q/2 \\ 2 & w.p. \ \ q/2 \end{cases}. \tag{2}$$

Solution: Let's first recognize that without the constraint that $P(W \neq 0) \leq q$, the entropy is maximized with a uniform distribution, just like we learned in class,

$$W_q = \begin{cases} 0 & w.p. \ \ \frac{1}{3} \\ 1 & w.p. \ \ \frac{1}{3} \\ 2 & w.p. \ \ \frac{1}{3} \end{cases}$$

When $q \geq \frac{2}{3}$, this uniform distribution is achievable, and is the distribution which maximizes the entropy.

What about when $q < \frac{2}{3}$? Let's consider $P(W \neq 0) = s$. The entropy is then

$$H(W) = -p \log_2(p) - (s-p) \log_2(s-p) - (1-s) \log_2(1-s)$$

Where $p = P(W = 1)$, $s - p = P(W = 2)$, and $1 - s = P(W = 0)$. Notice that the last term doesn't depend on p, so maximizing this with respect to p is the same as maximizing h_2 with a maximum possible value of q instead of 1. ($\frac{\partial}{\partial p} H(W) = \log_2(\frac{s}{p} - 1)$) Therefore, the maximum is obtained at $p^* = \frac{s}{2}$. Intuitively, this should make sense as we want the probabilities to be as uniformly distributed as possible.

But, should $s = q$ or should $s < q$? Substituting $p^* = \frac{s}{2}$ into $H(W)$ yields,

$$\begin{aligned} H(W) &= -\frac{s}{2} \log_2 \frac{s}{2} - \frac{s}{2} \log_2 \frac{s}{2} - (1-s) \log_2(1-s) \\ &= -s \log_2 \frac{s}{2} - (1-s) \log_2(1-s) \end{aligned}$$

Taking the derivative with respect to s:

$$\frac{\partial}{\partial s} H(W) = \log_2(1-s) - \log_2 \frac{s}{2}$$

$$= \log_2(\frac{2}{s} - 2)$$

Notice that this derivative is positive for $s < \frac{2}{3}$, so whatever value q takes on less than $\frac{2}{3}$, $H(W)$ will increase with s up to q. So the maximal entropy is achieved when

$$W_q = \begin{cases} 0 & w.p. \quad 1-q \\ 1 & w.p. \quad q/2 \\ 2 & w.p. \quad q/2 \end{cases}. \tag{3}$$

as expected from the hint.

Plotting $\phi(q)$ looks as follows:

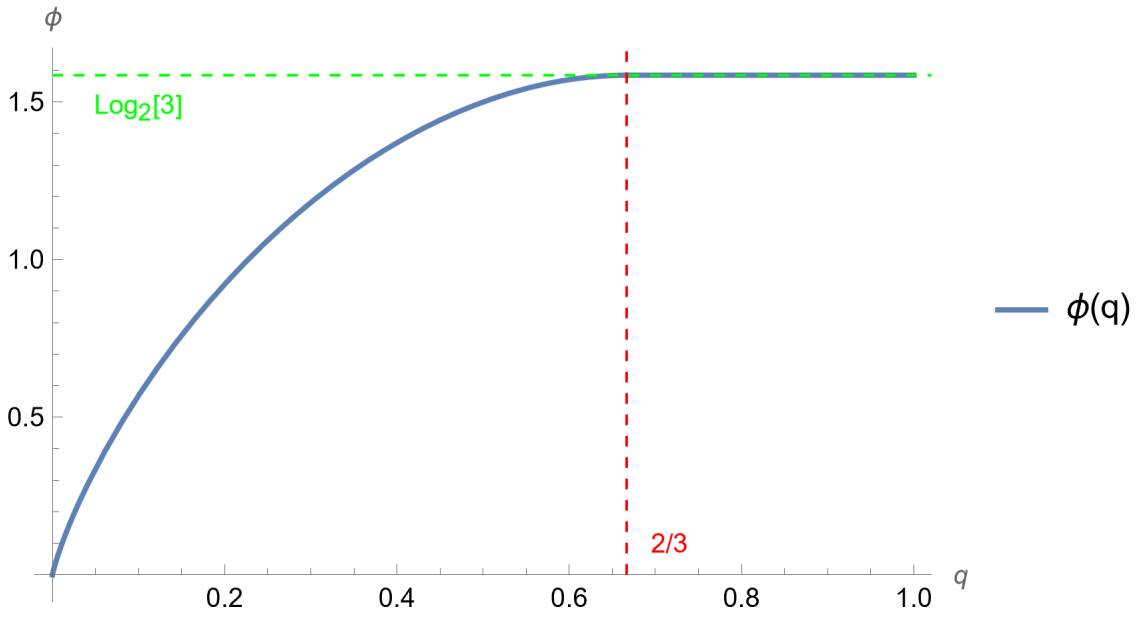

Figure 1: $\phi(q)$ for $0 \leq q \leq 1$.

(b) Consider the rate distortion function $R(D)$ of a ternary memoryless source U under Hamming distortion, i.e. $\mathcal{U} = \mathcal{V} = \{0, 1, 2\}$ and

$$d(u, v) = \begin{cases} 0 & if \quad u = v \\ 1 & if \quad u \neq v \end{cases}$$

for all $u \in \mathcal{U}, v \in \mathcal{V}$. Show that for $D \geq 0$:

$$R(D) \geq H(U) - \phi(D).$$

Solution: For $\ominus$ defined as subtraction modulo 3, we have

$I(U;V) = H(U) - H(U|V)$ Definition of mutual information

$= H(U) - H(U \ominus V|V)$ From invariance of entropy to translation of the RV (or to any one-to-one transformation)

$\geq H(U) - H(U \ominus V)$ Conditioning reduces entropy

$\geq H(U) - \phi(D)$ Due to $P(U \ominus V \neq 0) = E[d(U,V)] \leq D$ and the definition of ϕ.

Thus, $H(U) - \phi(D)$ lower bounds any mutual information in the feasible set over which the minimum in the definition of $R(D)$ is taken, and therefore lower bounds $R(D)$.

(c) Show that in the setting of the previous part, when U is the uniform ternary source

$$P(U = 0) = P(U = 1) = P(U = 2) = 1/3, \tag{4}$$

the lower bound is achieved with equality, i.e. $R(D) = \log 3 - \phi(D)$ for $D \geq 0$.

Solution: We need to find a distribution on (U, V) such that:

- U is uniform
- $U \ominus V$ is independent on V
- $U \ominus V \sim W_{q=D}$

Taking V to be uniform, and $U = V \oplus W_{q=D}$, for $W_{q=D}$ independent of V, satisfies these three conditions.

(d) Recall the random variable W_q defined in (4a). Consider now communication over a memoryless ternary-input ternary-output channel with mod-3 additive noise distributed as W_q, for some $0 \leq q \leq 2/3$. I.e., $\mathcal{X} = \mathcal{Y} = \{0, 1, 2\}$, and the relationship between the channel input and output is given by

$$Y = (X + W_q) \mod 3, \tag{5}$$

where X and W_q are independent. Show that the capacity of this channel is $C = \log 3 - \phi(q)$.

Solution:

$$\begin{aligned} I(X;Y) &\overset{(i)}{=} H(Y) - H(Y \mid X) \\ &\overset{(ii)}{=} H(Y) - H(Y \ominus X \mid X) \\ &\overset{(iii)}{=} H(Y) - H(W_q \mid X) \\ &\overset{(iv)}{=} H(Y) - H(W_q) \\ &\overset{(v)}{\leq} \log 3 - H(W_q) \\ &\overset{(vi)}{=} \log 3 - \phi(q) \end{aligned} \tag{6}$$

where (i) follows from the definition of mutual information, (ii) is due to invariance of entropy to translation of the RV (or to any one-to-one transformation), (iii) is

due to the channel model, (iv) to the independence of the additive channel noise component on the channel input, and (v) is because Y is ternary. Finally, (vi) follows from the fact that W_q is the distribution whose entropy achieves $\phi(q)$.

(e) Consider now a joint-source-channel-coding (JSCC) scenario where the uniform source defined in (4) is to be communicated over the channel defined in (5). What is the maximum achievable communication rate if the source is to be communicated losslessly?

Solution: The Source-Channel Separation Theorem tells us that a rate-distortion pair (ρ, D) is achievable iff

$$\rho \cdot R(D) \leq C$$

So, the maximum achievable rate of communication is

$$\begin{aligned}\rho \leq \frac{C}{R(D)} &= \frac{\log_2 3 - \phi(q)}{\log_2 3 - \phi(D)} \\ &= \frac{\log_2 3 - \phi(q)}{\log_2 3 - \phi(0)} = 1 - \frac{\phi(q)}{\log_2 3}\end{aligned}$$

(f) Consider again the JSCC scenario as in the previous part. What is the minimum achievable distortion for communication at rate $\rho = 1$ source symbols per channel use?

Solution: The Source-Channel Separation Theorem tells us that a rate-distortion pair (ρ, D) is achievable iff

$$\rho \cdot R(D) \leq C$$

For $\rho = 1$ and the $R(D)$ function found in part (c) coupled with the channel capacity found in part (d) yields,

$$\begin{aligned}\log_2 3 - \phi(D) &\leq \log_2 3 - \phi(q) \\ \phi(D) &\geq \phi(q)\end{aligned}$$

Since $\phi(x)$ is monotonically increasing for $0 \leq x \leq \frac{2}{3}$, The smallest D can be while satisfying the above condition is q. Therefore, the minimum achievable distortion is $D^* = q$.

(g) Suggest a concrete simple scheme for the setting of the previous part which achieves the optimum performance, that is, communicates at a rate of 1 source symbols per channel use and attains the minimum achievable distortion you found in the previous part. Explain your reasoning.

Solution: If we only recover Y as our reconstruction for X, we will have a probability of error, $p_{error} = P(W_q \neq 0) = q$, and an expected Hamming distortion $E[d(X, Y)] = 0 * (1 - q) + 1 * q = q$.

5. (*20 points*)

(a) Express the set

$$\left\{ u^n \in \{0,1\}^n : \frac{1}{n}\sum_{i=1}^{n} u_i \geq \gamma \right\}$$

as a union of types.

Solution: Let S be the set in question. Given u^n, we have

$$\frac{1}{n}\sum_{i=1}^{n} u_i \geq \gamma \Leftrightarrow p^{u^n}(1) \geq \gamma$$

where p^{u^n} is the empirical distribution of u^n. Hence,

$$S = \bigcup_{p \in \mathbb{P}^n : p(1) \geq \gamma} T(p).$$

(b) For $U_i \overset{i.i.d}{\sim} \texttt{Ber}(p)$, show that

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} U_i \geq \gamma \right) \doteq 2^{-nD\left(\texttt{Ber}(\gamma)\middle\|\texttt{Ber}(p)\right)} \quad \text{for any} \quad \gamma \in [p,1].$$

Hint: note that for any $\gamma \in [p,1]$

$$\min_{q \in [\gamma,1]} D\left(\texttt{Ber}(q)\middle\|\texttt{Ber}(p)\right) = D\left(\texttt{Ber}(\gamma)\middle\|\texttt{Ber}(p)\right).$$

Solution: By (a), the event of interest is given by

$$\mathbb{P}\left(\bigcup_{\widehat{p} \in S} T(p) \right) \leq |\mathbb{P}^n| \max_{\widehat{p} \in S} \mathbb{P}(T(\widehat{p})) \leq (n+1)2^{-n \min_{\widehat{p} \in S} D(\widehat{p}\|p)} \doteq 2^{-nD(\gamma\|p)}. \tag{7}$$

The lower bound is similarly obtained by noting that the sum of elements is lower bounded by the maximum value in the sum:

$$\mathbb{P}\left(\bigcup_{\widehat{p} \in S} T(p) \right) \geq \max_{\widehat{p} \in S} \mathbb{P}(T(\widehat{p})) \doteq 2^{-n \min_{\widehat{p} \in S} D(\widehat{p}\|p)} = 2^{-nD(\gamma\|p)}. \tag{8}$$

Now consider communicating one bit $X \sim \texttt{Ber}(1/2)$ via n uses of a BSC(p) with a repetition code. The ith channel output is thus given by

$$Y_i = X \oplus Z_i,$$

where $\{Z_i\}_{i\geq 1}$ are IID $\sim \texttt{Ber}(p)$, independent of X. Assume in what follows that $p < \frac{1}{2}$.

(c) What is the optimal decoding rule in the sense of minimizing the probability of error?
Solution: Since $\mathbb{P}(X_i = 1) = \frac{1}{2}$ and $p < \frac{1}{2}$, the optimal rule to minimize the probability of error is a majority vote corresponding to the maximum likelihood estimate of X.

(d) Let $P_e(n, p)$ be the probability of error of the optimal decoder from the previous part. Find the exponential decay rate of $P_e(n, p)$, i.e., what is

$$\lim_{n\to\infty} -\frac{1}{n} \log P_e(n, p)?$$

Solution:
Note that

$$\begin{aligned}
\mathbb{P}_e(n,p) &= \frac{1}{2}\mathbb{P}(\widehat{X} = 1|X = 0) + \frac{1}{2}\mathbb{P}(\widehat{X} = 0|X = 1) && (9)\\
&= \frac{1}{2}\mathbb{P}_p\left(\frac{1}{n}\sum_i Y_i \geq \frac{1}{2}\right) + +\frac{1}{2}\mathbb{P}_{1-p}\left(\frac{1}{n}\sum_i Y_i \leq \frac{1}{2}\right) && (10)\\
&= \frac{1}{2}\mathbb{P}_p\left(\frac{1}{n}\sum_i Y_i \geq \frac{1}{2}\right) + \frac{1}{2}\mathbb{P}_{1-p}\left(\frac{1}{n}\sum_i (1 - Y_i) \leq \frac{1}{2}\right) && (11)\\
&= \mathbb{P}_p\left(\frac{1}{n}\sum_i Y_i \geq \frac{1}{2}\right) && (12)\\
&\doteq 2^{-nD(\text{Ber}(1/2)\|\text{Ber}(p))}. && (13)
\end{aligned}$$

So

$$\lim_{n\to\infty} -\frac{1}{n} \log P_e(n, p) = D(\texttt{Ber}(1/2)\|\texttt{Ber}(p)).$$