EE 276 - Information Theory
Midterm
Feb 16, 2024

1. There are a total of 5 questions: 3 shorter questions, and 2 longer ones. You have 2 hours to take the exam. Questions have different numbers of points as indicated before each sub-problem. There are a total of 100 points, with 2 additional bonus points.

2. Please write all answers in the designated area underneath the question. If you need more room for your answer, please indicate as such 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. All answers should be justified, unless otherwise stated.

5. The exam is closed book but you are allowed one double-sided sheet of notes. No other materials are allowed.

6. Calculators are not allowed.

Good luck!

Name:

SUID:

1. Let X and Y be random variables with finite alphabet and non-zero entropy. Consider the quantity

$$\rho := \frac{I(X;Y)}{\min\{H(X), H(Y)\}}.$$

   (a) *(8 points)* Show that $0 \leq \rho \leq 1$.

   Solution: The quantity ρ is symmetric. So without loss of generality, we can assume that $H(X) \leq H(Y)$ and hence

$$\rho = \frac{I(X;Y)}{\min\{H(X), H(Y)\}} = \frac{H(X) - H(X|Y)}{H(X)} = 1 - \frac{H(X|Y)}{H(X)}.$$

   Since conditioning reduces entropy and entropy is nonnegative, we have

$$0 \leq \frac{H(X|Y)}{H(X)} \leq 1,$$

   and hence $0 \leq \rho \leq 1$.

   (b) *(8 points)* Interpret the extreme cases of $\rho = 0$ and $\rho = 1$: what can you say about the relationship between X and Y under each of these cases?

   Solution: $\rho = 0$ if and only if $I(X;Y) = 0$ which holds if and only if X, Y are independent. Meanwhile, $\rho = 1$ holds if $I(X;Y) = H(X)$ ($I(X;Y) = H(Y)$) and hence $H(X|Y) = 0$ (or $H(Y|X) = 0$) implying that one is a deterministic function of the other.

2. Let the discrete random variables X, Y, Z be such that

$$Z = X + Y.$$

(a) *(8 points)* Show that $H(Z) \leq H(X) + H(Y)$.

Solution: We have

$$H(X + Y) \leq H(X, Y) = H(X) + H(Y|X) \leq H(X) + H(Y).$$

The first inequality holds since $X + Y$ is a deterministic function of (X, Y), and the second holds because conditioning reduces entropy.

(b) *(6 points)* Under what conditions does the inequality hold with equality?

Solution: The first inequality above holds with equality when the map from (X, Y) to $Z = X + Y$ is invertible, meaning that the values of X, Y can be determined from the sum. One example of this is if $X \in \{0, 1\}$ with equal probability and $Y \in \{2, 4\}$ with equal probability. Meanwhile, the second inequality holds with equality if X and Y are independent.

3. *(10 points)* A memoryless source $X_1, X_2, \ldots$ is drawn from $\mathcal{X} = \{a, b, c\}$ with respective probabilities $(1/2, 1/4, 1/4)$. Consider the typical set $\mathcal{A}_\epsilon^{(n)}$ for $n = 8$ and $\epsilon = 1/10$. How many occurrences of the letter "a" are there in a sequence belonging to this set? (i.e., does this set contain sequences with one "a", two "a"s, ..., etc.?). *You don't need to perform long and difficult arithmetic computations to answer this question.*

Solution: A sequence is typical (i.e. in the typical set, $\mathcal{A}_\epsilon^{(n)}$) if for $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ we have,

$$2^{-n(H(X)+\epsilon)} \leq p(x_1, x_2, \ldots, x_n) \leq 2^{-n(H(X)-\epsilon)}$$

Note that for the above source,

$$\begin{aligned} H(X) &= -\left(\frac{1}{2}\log\frac{1}{2} + \frac{1}{4}\log\frac{1}{4} + \frac{1}{4}\log\frac{1}{4}\right) \\ &= \left(\frac{1}{2}\log 2 + \frac{1}{4}\log 4 + \frac{1}{4}\log 4\right) \\ &= \left(\frac{1}{2} + \frac{1}{2} + \frac{1}{2}\right) = \frac{3}{2} \end{aligned}$$

for $n = 8$ and $\epsilon = 0.1$, we have

$$\begin{aligned} 2^{-8(\frac{3}{2}+\frac{1}{10})} &\leq p(x_1, x_2, \ldots, x_n) \leq 2^{-8(\frac{3}{2}-\frac{1}{10})} \\ 2^{-\frac{64}{5}} &\leq p(x_1, x_2, \ldots, x_n) \leq 2^{-\frac{56}{5}} \\ 2^{-12\frac{4}{5}} &\leq p(x_1, x_2, \ldots, x_n) \leq 2^{-11\frac{1}{5}} \end{aligned}$$

The probabilities for a sequence with n_a a's, n_b b's, and n_c c's $(n_a + n_b + n_c = n)$ can be written as $p(x_1, x_2, \ldots, x_n) = \left(\frac{1}{2}\right)^{n_a}\left(\frac{1}{4}\right)^{n_b}\left(\frac{1}{4}\right)^{n_c} = \left(\frac{1}{2}\right)^{n_a}\left(\frac{1}{4}\right)^{n-n_a} = 2^{-(n_a+2(n-n_a))} = 2^{-(2n-n_a)}$. We can see trivially that this exponent will be a whole number. The only whole number between $-11\frac{1}{5}$ and $-12\frac{4}{5}$ is -12, so setting $2n - n_a = 12$ and $n = 8$ yields $n_a = 4$. So, the sequences which lie in typical set are those which have exactly 4 a's in them.

If you are curious how large $\mathcal{A}_\epsilon^{(n)}$ is, there are $\binom{n}{n_a}$ ways of choosing where the n_a a's go, then, of the remaining $n - n_a$ spaces available, there are $\sum_{i=0}^{n-n_a}\binom{n-n_a}{i}$ possible ways to arrange the b's and c's. So, the total size of $\mathcal{A}_\epsilon^{(n)}$ is

$$|\mathcal{A}_\epsilon^{(n)}| = \binom{n}{n_a}\sum_{i=0}^{n-n_a}\binom{n-n_a}{i}$$

For the above example, we have

$$\begin{aligned} |\mathcal{A}_\epsilon^{(n)}| &= \binom{8}{4}\sum_{i=0}^{4}\binom{4}{i} = \binom{8}{4}2^4 \\ &= 70 \cdot 16 = 1120 \end{aligned}$$

Since every sequence in the typical set has the probability 2^{-12}, the probability of a sequence being in the typical set is just

$$P(X^n \in \mathcal{A}_\epsilon^{(n)}) = 2^{-12} \cdot 1120 \approx 0.2734375$$

4. (*Each part of this problem can be attempted independently of the other parts, using only what is stated in the prompts of the preceding parts.*)

In this problem, we will study the entropy of the English language as Shannon did. In what follows, we will model an infinitely long string of English text as a random process

$$X_1, X_2, X_3 \dots$$

where X_i takes values in $\mathcal{X} := \{a, \dots, z\}, |\mathcal{X}| = 26$. Here, X_i represents the ith letter in a long English string. Note that the X_is are *not independent*, but for simplicity, we'll assume that the process is *stationary*, meaning that

$$p(x_1, \dots, x_n) = p(x_{1+k}, \dots, x_{n+k})$$

for all $n, k \geq 0$ and symbols $x_j \in \mathcal{X}$.

For $n \geq 1$, let us define the *n-gram entropy* as

$$\begin{aligned} \mathcal{H}_1 &:= H(X_1) \\ \mathcal{H}_n &:= H(X_n | X_{n-1}, X_{n-2}, \dots, X_1), \quad \text{for} \quad n \geq 2, \end{aligned}$$

which can be thought of as the entropy of the next letter in position n given the previous $n-1$ letters. A natural definition of the entropy of English is then

$$\mathcal{H}_{\text{Eng}} := \lim_{n \to \infty} \mathcal{H}_n.$$

(a) *(8 points)* Show that $\mathcal{H}_n \geq \mathcal{H}_{n+1}$.

Solution: The process is stationary. So, we should have $H(X_n | X_{n-1} X_{n-2} \dots X_1) = H(X_{n+1} | X_n X_{n-1} \dots X_2) = \mathcal{H}_n$. Since conditioning reduces entropy, we must have $\mathcal{H}_{n+1} = H(X_{n+1} | X_n X_{n-1} \dots X_1) \leq \mathcal{H}_n$.

(b) *(4 points)* Using results from class, or otherwise, argue that $\mathcal{H}_n$ converges, i.e., that $\mathcal{H}_{\text{Eng}}$ exists and is finite.

Solution: We learned in class that this limit exists as long as the process is stationary. This can also be deduced from the previous part: namely, we know that $\mathcal{H}_0 \geq \mathcal{H}_n$. Furthermore, we know that conditional entropy is non-negative, so $\mathcal{H}_n \geq 0$. Due to the monotone convergence theorem, since $\mathcal{H}_n$ is bounded and monotonically decreasing, the limit exists.

For the sake of simplicity, in what follows, we'll only consider approximating $\mathcal{H}_{\text{Eng}}$ using $n = 1$ and $n = 2$.

(c) *(6 points)* Show that the maximum values that $\mathcal{H}_1$ and $\mathcal{H}_2$ can take on are $\mathcal{H}_1^\star$ and $\mathcal{H}_2^\star$, respectively, which are given by

$$\mathcal{H}_1^\star = \mathcal{H}_2^\star = \log 26 \ \ (\approx 4.7 \text{ bits}).$$

Solution: We have $\mathcal{H}_1 = H(X_1)$ and $\mathcal{H}_2 = H(X_2|X_1) = H(X_2, X_1) - H(X_1)$

Uniform distributions maximize entropy, so we want $p(i) = \frac{1}{N}$ $\forall i$ and $q(i, j) = \frac{1}{N^2}$ $\forall (i, j)$ This gives:

$$\mathcal{H}_1 = \log N$$
$$\mathcal{H}_2 = \log N^2 - \log N = \log N$$

For $N = 26$, we have $\mathcal{H}_1 = \mathcal{H}_2 = \log 26 = 4.7$

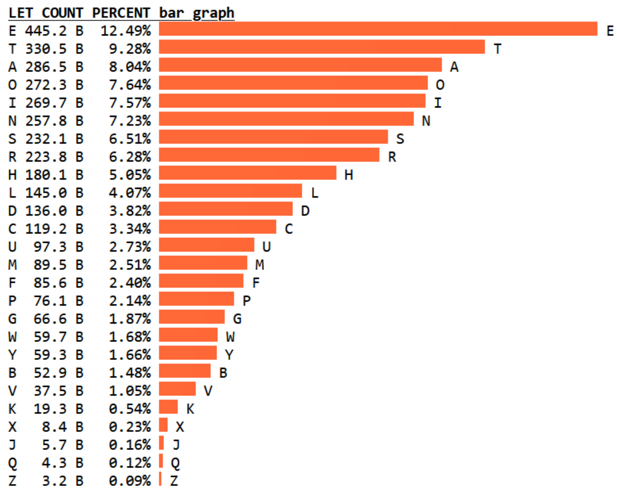

Figure 1: Letter Frequencies calculated by Peter Norvig in 2012.

When Shannon and his wife Mary estimated $\mathcal{H}_1$ and $\mathcal{H}_2$ from English text, they found that

$$\mathcal{H}_1 \approx 4.14 \text{ bits} \quad \text{and} \quad \mathcal{H}_2 \approx 3.56 \text{ bits}.$$

(d) *(5 points)* Give a reason why $\mathcal{H}_1$ for English is lower than $\mathcal{H}_1^\star$.

Solution: English letters are not distributed normally! On HW1, we saw that one-hot probability vectors across the letters minimizes entropy. The distribution of English is between these two extremes and is show in figure 1.

(e) *(5 points)* In general, you showed in part (a) that $\mathcal{H}_2 \leq \mathcal{H}_1$. What does the fact that $\mathcal{H}_2$ is considerably smaller than $\mathcal{H}_1$ tell us about letter pairs in English?

Solution: $\mathcal{H}_2$ is strictly less than $\mathcal{H}_1$ because English is contextual, English letters depend on what comes before them, and therefore not independent. This means that conditioning on the letters we have seen in the past, will strictly reduce the entropy. You can see in figure 2 the empirical joint distribution of bigrams. There are only seven bigrams that do not occur among the 2.8 trillion mentions: JQ, QG, QK, QY, QZ, WQ, and WZ. They are shown as crossed out.

Figure 2: Digram frequencies computed by Peter Norvig in 2012. There are only seven bigrams that do not occur among the 2.8 trillion instances of digrams: JQ, QG, QK, QY, QZ, WQ, and WZ. They are shown as crossed out.

5. (*Each part of this problem can be attempted independently of the other parts, using only what is stated in the prompts of the preceding parts.*)

Consider a channel given by

$$Y = X + Z \tag{1}$$

where the input X, output Y and noise Z are nonnegative. Furthermore, Z and X are independent. Let Z have an exponential distribution with parameter $\lambda > 0$, which we denote by writing $Z \sim \texttt{exp}(\lambda)$. More precisely, this means that the density of Z is given by

$$f(z) = \begin{cases} 0 & z < 0 \\ \lambda e^{-\lambda z} & z \geq 0. \end{cases}$$

It may also be useful to recall that this implies that $\mathbb{E}[Z] = 1/\lambda$.

In this problem, you will find the capacity of this channel, under the constraint on the input

$$\mathbb{E}[X] \leq P. \tag{2}$$

(a) *(8 points)* Show that the differential entropy of Z is given by

$$h(Z) = -\log \lambda + \frac{1}{\ln 2}.$$

Solution:

$$\begin{aligned} h(Z) &= -\mathbb{E}[\log(\lambda e^{-\lambda Z})] \\ &= -\log(\lambda) + \lambda \log(e) \mathbb{E}[Z] \\ &= -\log(\lambda) + \frac{1}{\ln 2} \end{aligned}$$

where we used the given value of the expectation of Z.

(b) *(8 points)* Show that for any random variable $W \geq 0$ with $\mathbb{E}[W] \leq M$, we have

$$h(W) \leq -\log \frac{1}{M} + \frac{1}{\ln 2}.$$

Solution: The proof here follows the proof for the Gaussian from lecture, or the distribution Ce^{-x^4} from the homework. First, observe that the bound in question corresponds to the differential entropy of an exponential, random variable with parameter $1/M$. So for a given random variable W with distribution $\mathbb{P}_W$ that satisfies the expectation constraint given, we have

$$\begin{aligned} 0 &\leq D(\mathbb{P}_W \| \texttt{exp}(1/M)) \\ &= -h(W) - \mathbb{E}[\log(M^{-1}e^{-W/M})] \\ &= -h(W) + \log M + M^{-1}\log(e)\mathbb{E}[W]\cdot \\ &\leq -h(W) + \log M + \frac{1}{\ln 2}, \end{aligned}$$

The result follows by adding $h(W)$ to both sides.

(c) *(8 points)* Use the claims of previous two parts to show that

$$I(X;Y) \leq \log(1 + P\lambda)$$

for Y as in (1) under the constraint (2) on X.

Solution: We have

$$I(X;Y) = h(X+Z) - h(X+Z|X) = h(X+Z) - h(Z)$$

since X and Z are independent. Now,

$$h(Z) = -\log(\lambda) + \frac{1}{\ln 2}$$

from part (a). Meanwhile, to bound $h(X+Z)$, we note that

$$\mathbb{E}[X+Z] = \mathbb{E}[X] + \mathbb{E}[Z] \leq P + \frac{1}{\lambda},$$

so we can apply part (b) to conclude

$$h(X+Z) \leq -\log \frac{1}{P + 1/\lambda} + \frac{1}{\ln 2}.$$

So

$$I(X;Y) \leq \log\left(P + \frac{1}{\lambda}\right) + \log(\lambda) = \log(1 + P\lambda)$$

In what follows, you may use the following fact: *Let* $U \sim \texttt{exp}\,(\lambda/(1+P\lambda))$. *If* X *is the random variable defined as*

$$X = \begin{cases} 0 & \text{with probability } \frac{1}{1+P\lambda} \\ U & \text{with probability } \frac{P\lambda}{1+P\lambda} \end{cases}, \tag{3}$$

then $Y = X + Z$ *has the distribution* $\texttt{exp}(\lambda/(1+P\lambda))$.

(d) *(8 points)* Use this fact with the claims of the previous parts to find the capacity of the channel in (1) under the constraint given in (2).

Solution: Consider X with the distribution in (3). We have

$$\mathbb{E}[X] = \frac{P\lambda}{1+P\lambda}\frac{1+P\lambda}{\lambda} = P.$$

Hence, such an X is a valid input to our channel. Since Y then has the distribution $\texttt{exp}(\lambda/(1+P\lambda))$, we have for this given input X,

$$\begin{aligned} I(X;Y) &= h(Y) - h(Y|X) \\ &= h(Y) - h(Z) \\ &= -\log\frac{\lambda}{1+P\lambda} + \log\lambda \\ &= \log(1+P\lambda). \end{aligned}$$

Since this is the maximum that $I(X;Y)$ can be under our constraint by part (c), then this must be the capacity of the channel.

(e) **(Bonus 2 points)** Prove the fact above about the distribution of Y. Namely, show that if X follows the distribution given in (3), then Y has the distribution $\texttt{exp}(\lambda/(1+P\lambda))$. *You may use that the moment generating function of an exponential random variable with parameter* λ *is given by* $\phi(t) = \lambda/(\lambda - t) \quad \text{for} \quad t < \lambda$.