EE276: Homework #1 Solutions

Due by 11:59pm PT, Tuesday, 26 Jan 2021

Please submit your solutions to Gradescope.

1. **Example of Entropy & Joint Entropy.**
Let $p(x, y)$ be given by

Y X	0	1
0	$\frac{1}{4}$	$\frac{1}{4}$
1	0	$\frac{1}{2}$

Find

(a) $H(X), H(Y)$.

(b) $H(X|Y = 0), H(X|Y = 1), H(Y|X = 0), H(Y|X = 1)$.

(c) $H(X|Y), H(Y|X)$.

(d) $H(X, Y)$.

(e) $I(X; Y)$.

(f) Compare the quantities in part (b) to the entropy values found in part (a). Does any of these violate the fact that conditioning reduces entropy ($H(X|Y) \leq H(X)$) as discussed in class? Comment very briefly.

Solution: Example of joint entropy

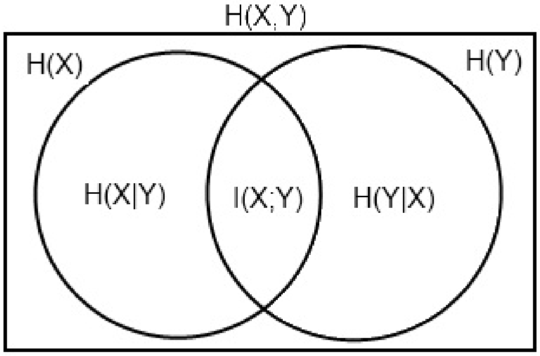


Figure 1: Venn diagram to illustrate the relationships of entropy and relative entropy

(a) $H(X) = \frac{1}{2} \log \frac{2}{1} + \frac{1}{2} \log \frac{2}{1} = 1.0$ bits.
$= H(Y) = \frac{1}{4} \log 4 + \frac{3}{4} \log \frac{4}{3} = 0.811$ bits.

(b)

$$\begin{aligned}H(X|Y=0) =& p(x=0|y=0)\log\frac{1}{p(x=0|y=0)} + p(x=1|y=0)\log\frac{1}{p(x=1|y=0)}\\ &= 1\log\frac{1}{1} + 0\log\frac{1}{0}\\ &= 0 \text{ bits}\end{aligned}$$

$$\begin{aligned}H(X|Y=1) =& p(x=0|y=1)\log\frac{1}{p(x=0|y=1)} + p(x=1|y=1)\log\frac{1}{p(x=1|y=1)}\\ &= \frac{1}{3}\log\frac{3}{1} + \frac{2}{3}\log\frac{3}{2}\\ &= 0.9183 \text{ bits}\end{aligned}$$

$$\begin{aligned}H(Y|X=0) =& p(y=0|x=0)\log\frac{1}{p(y=0|x=0)} + p(y=1|x=0)\log\frac{1}{p(y=1|x=0)}\\ &= \frac{1}{2}\log\frac{2}{1} + \frac{1}{2}\log\frac{2}{1}\\ &= 1 \text{ bit}\end{aligned}$$

$$\begin{aligned}H(Y|X=1) =& p(y=0|x=1)\log\frac{1}{p(y=0|x=1)} + p(y=1|x=1)\log\frac{1}{p(y=1|x=1)}\\ &= 0\log\frac{1}{0} + 1\log\frac{1}{1}\\ &= 0 \text{ bits}\end{aligned}$$

(c) $H(X|Y) = \frac{1}{4}H(X|Y=0) + \frac{3}{4}H(X|Y=1) = 0.689$ bits.
$H(Y|X) = \frac{1}{2}H(Y|X=0) + \frac{1}{2}H(Y|X=1) = 0.5$ bits.

(d) $H(X,Y) = \frac{1}{4}\log 4 + \frac{1}{4}\log 4 + \frac{1}{2}\log 2 = 1.5$ bits.

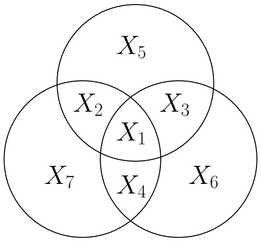
(e) $I(X;Y) = H(Y) - H(Y|X) = 0.311$ bits.

We can draw a Venn diagram for the quantities in (a) through (e). See Figure 1.

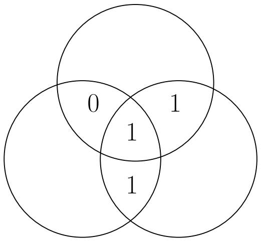
(f) We observe that $0 = H(Y|X=1) < H(Y|X) = 0.5$ and that $1 = H(Y|X=0) > H(Y|X) = 0.5$. This does not violate the fact that conditioning reduces entropy because that fact holds on average saying that $H(Y|X) < H(Y)$. This does not preclude values x of X for which $H(Y|X=x) > H(Y)$. An example from Cover and Thomas: “For example, in a court case, specific new evidence might increase uncertainty, but on the average evidence decreases uncertainty.”

2. **Entropy of Hamming Code.**
Hamming code is a simple error-correcting code that can correct up to one error in a sequence of bits. Now consider information bits $X_1, X_2, X_3, X_4 \in \{0,1\}$ chosen

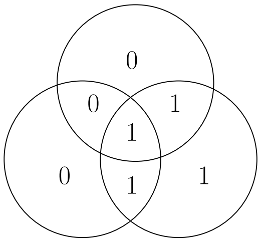
uniformly at random, together with check bits X_5, X_6, X_7 chosen to make the parity of the circles even.
(eg: $X_1 + X_2 + X_4 + X_7 = 0 \mod 2$)



Thus, for example,



becomes



That is, 1011 becomes 1011010.

(a) What is the entropy of $H(X_1, X_2, ..., X_7)$?

Now we make an error (or not) in one of the bits (or none). Let $\mathbf{Y} = \mathbf{X} \oplus \mathbf{e}$, where $\mathbf{e}$ is equally likely to be $(1, 0, \ldots, 0), (0, 1, 0, \ldots, 0), \ldots, (0, 0, \ldots, 0, 1)$, or $(0, 0, \ldots, 0)$, and $\mathbf{e}$ is independent of $\mathbf{X}$.

(b) Show that one can recover the message $\mathbf{X}$ perfectly from $\mathbf{Y}$. (Please provide a justification, detailed proof not required.)

(c) What is $H(\mathbf{X}|\mathbf{Y})$?

(d) What is $I(\mathbf{X}; \mathbf{Y})$?

(e) What is the entropy of $\mathbf{Y}$?

Solution: Entropy of Hamming Code.

(a) By the chain rule,

$$H(X_1, X_2, ..., X_7) = H(X_1, X_2, X_3, X_4) + H(X_5, X_6, X_7|X_1, X_2, X_3, X_4).$$

Since X_5, X_6, X_7 are all deterministic functions of X_1, X_2, X_3, X_4, we have

$$H(X_5, X_6, X_7|X_1, X_2, X_3, X_4) = 0.$$

And since X_1, X_2, X_3, X_4 are independent Bernoulli(1/2) random variables,

$$H(X_1, X_2, ..., X_7) = H(X_1) + H(X_2) + H(X_3) + H(X_4) = 4.$$

(b) We first note that the Hamming code can detect one error. This follows from the fact that a flip of a single bit will result in the parity of at least one of the circles getting odd. Now, depending on which parities become odd, one can detect the precise location of the error. For example, if all three parities are odd, then X_1 is received in error. Similarly, if only the top circle parity is odd, X_5 is in error. In a similar manner, one can verify that $\mathbf{X}$ can be recoverd from $\mathbf{Y}$.

(c) As shown in (b), $\mathbf{X}$ is a deterministic function of $\mathbf{X} \oplus \mathbf{e}$. So $H(\mathbf{X}|\mathbf{Y}) = 0$.

(d) $I(\mathbf{X};\mathbf{Y}) = H(\mathbf{X}) - H(\mathbf{X}|\mathbf{Y}) = H(\mathbf{X}) = 4$.

(e) We will expand $H(\mathbf{X} \oplus \mathbf{e}, \mathbf{X})$ in two different ways, using the chain rule. On one hand, we can write

$$\begin{aligned} H(\mathbf{X} \oplus \mathbf{e}, \mathbf{X}) &= H(\mathbf{X} \oplus \mathbf{e}) + H(\mathbf{X}|\mathbf{X} \oplus \mathbf{e}) \\ &= H(\mathbf{X} \oplus \mathbf{e}). \end{aligned}$$

In the last step, $H(\mathbf{X}|\mathbf{X} \oplus \mathbf{e}) = 0$ because $\mathbf{X}$ is a deterministic function of $\mathbf{X} \oplus \mathbf{e}$. On the other hand, we can also expand $H(\mathbf{X} \oplus \mathbf{e}, \mathbf{X})$ as follows:

$$\begin{aligned} H(\mathbf{X} \oplus \mathbf{e}, \mathbf{X}) &= H(\mathbf{X}) + H(\mathbf{X} \oplus \mathbf{e}|\mathbf{X}) \\ &= H(\mathbf{X}) + H(\mathbf{X} \oplus \mathbf{e} \oplus \mathbf{X}|\mathbf{X}) \\ &= H(\mathbf{X}) + H(\mathbf{e}|\mathbf{X}) \\ &= H(\mathbf{X}) + H(\mathbf{e}) \\ &= 4 + H(\mathbf{e}) \\ &= 4 + \log_2 8 \\ &= 7. \end{aligned}$$

The second equality follows since XORing with $\mathbf{X}$ is a one-to-one deterministic function (when conditioned on $\mathbf{X}$). The third equality follows from the well-known property of XOR that $y \oplus y = 0$. The fourth equality follows since the error vector

$\mathbf{e}$ is independent of $\mathbf{X}$. The fifth equality follows since from part (a), we know that $H(\mathbf{X}) = 4$. The sixth equality follows since $\mathbf{e}$ is uniformly distributed over eight possible values: either there is an error in one of seven positions, or no error at all.

Equating our two different expansions for $H(\mathbf{X} \oplus \mathbf{e}, \mathbf{X})$, we have

$$H(\mathbf{X} \oplus \mathbf{e}, \mathbf{X}) = H(\mathbf{X} \oplus \mathbf{e}) = 7.$$

The entropy of $\mathbf{Y} = \mathbf{X} \oplus \mathbf{e}$ is 7 bits.

This result is closely related to the fact that the code in consideration, the Hamming [7,4,3] code, is a perfect code (`https://en.wikipedia.org/wiki/Hamming_bound#Perfect_codes`), and hence $\mathbf{X} \oplus \mathbf{e}$ is uniformly distributed in $\{0,1\}^7$.

3. **Entropy of functions of a random variable.**
Let X be a discrete random variable. Show that the entropy of a function of X is less than or equal to the entropy of X by justifying the following steps:

$$\begin{aligned} H(X, g(X)) &\overset{(a)}{=} H(X) + H(g(X)|X) \\ &\overset{(b)}{=} H(X). \\ H(X, g(X)) &\overset{(c)}{=} H(g(X)) + H(X|g(X)) \\ &\overset{(d)}{\geq} H(g(X)). \end{aligned}$$

Thus $H(g(X)) \leq H(X)$.

Solution: Entropy of functions of a random variable.

(a) $H(X, g(X)) = H(X) + H(g(X)|X)$ by the chain rule for entropies.

(b) $H(g(X)|X) = 0$, since for any particular value of X, g(X) is deterministic, and hence $H(g(X)|X) = \sum_x p(x) H(g(X)|X = x) = \sum_x 0 = 0$.

(c) $H(X, g(X)) = H(g(X)) + H(X|g(X))$ again by the chain rule.

(d) $H(X|g(X)) \geq 0$, with equality iff X is a function of $g(X)$, i.e., $g(\cdot)$ is one-to-one. Hence $H(X, g(X)) \geq H(g(X))$.

Combining parts (b) and (d), we obtain $H(X) \geq H(g(X))$.

4. **Data Processing Inequality.**
The random variables X, Y and Z form a Markov triplet $(X - Y - Z)$ if $p(z|y) = p(z|y, x)$, and as a corollary $p(x|y) = p(x|y, z)$. If X, Y, Z form a Markov triplet $(X - Y - Z)$, show that:

(a) $H(X|Y) = H(X|Y, Z)$ and $H(Z|Y) = H(Z|X, Y)$

(b) $H(X|Y) \leq H(X|Z)$

(c) $I(X;Y) \geq I(X;Z)$ and $I(Y;Z) \geq I(X;Z)$

(d) $I(X;Z|Y) = 0$

The following definition may be useful:

Definition: The *conditional mutual information* of random variables X and Y given Z is defined by

$$\begin{aligned} I(X;Y|Z) &= H(X|Z) - H(X|Y,Z) \\ &= \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)} \end{aligned}$$

Solution: Data Processing Inequality.

(a)

$$\begin{aligned} H(X|Y) &= \sum_{x,y} -p(x,y) \log(p(x|y)) \\ &= \sum_{x,y,z} -p(x,y,z) \log(p(x|y)) \\ &= \sum_{x,y,z} -p(x,y,z) \log(p(x|y,z)) \\ &= H(X|Y,Z) \end{aligned}$$

where the third equality uses the fact that X and Z are conditionally independent given Y. A similar argument can be used to show $H(Z|Y) = H(Z|X,Y)$.

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

(c) $I(X;Y) = H(X) - H(X|Y) \geq H(X) - H(X|Z) = I(X;Z)$.

(d) We showed that $H(X|Y) = H(X|Z,Y)$, therefore, $I(X;Z|Y) = H(X|Y) - H(X|Z,Y) = 0$.

5. **Entropy of time to first success.**
A fair coin is flipped until the first head occurs. Let X denote the number of flips required.

(a) Find the entropy $H(X)$ in bits. The following expressions may be useful:

$$\sum_{n=1}^{\infty} r^n = r/(1-r), \quad \sum_{n=1}^{\infty} nr^n = r/(1-r)^2.$$

(b) To find the value of X, construct an "efficient" sequence of yes-no questions of the form, "Is X contained in the set S?". Compare $H(X)$ to the expected number of questions required to determine X.

(c) Let Y denote the number of flips until the second head appears. Thus, for example, $Y = 5$ if the second head appears on the 5th flip. Argue that $H(Y) = H(X_1 + X_2) < H(X_1, X_2) = 2H(X)$, and interpret in words. (Here, X_1 is the number of flips for the first head, and X_2 is the number of flips for the second head, after the occurence of the first head)

Solution: Entropy of time to first success

(a) The distribution of the number X of tosses until the first head appears is a geometric distribution with parameter $p = 1/2$: $P(X = n) = pq^{n-1}$ (where $q = 1 - p$), $n \in \{1, 2, \ldots\}$.

$$\begin{aligned} H(X) &= -\sum_{n=1}^{\infty} pq^{n-1} \log(pq^{n-1}) \\ &= -\left[\sum_{n=0}^{\infty} pq^n \log p + \sum_{n=0}^{\infty} npq^n \log q\right] \\ &= \frac{-p \log p}{1-q} - \frac{pq \log q}{p^2} \\ &= \frac{-p \log p - q \log q}{p} \\ &= H(p)/p \text{ bits.} \end{aligned}$$

If $p = 1/2$, then $H(X) = 2$ bits. Note also that this quantity, $H(p)/p$, makes perfect sense: the coin tosses are independent, so each coin toss gives us $H(p)$ bits of entropy. On average, we make $1/p$ tosses until we get the first head, so $H(p)/p$ should be the total entropy.

(b) This problem invites you to use your intuition to guess an answer. A good yes-or-no question is one whose entropy is as high as possible, that is, as close to one as possible. We model the question as a random variable Y that can take on two values: YES and NO, with probabilities p and $(1-p)$. The entropy of the question is then just $H(p)$. Since $I(X;Y) = H(Y) - H(Y|X) \leq H(Y)$, and the maximum entropy of Y is $H(1/2) = 1$, we can't learn more than 1-bit of information about X from any single question Y.

Now consider the sequence of questions, $Y_1 =$ "Is $X = 1$?", $Y_2 =$ "Is $X = 2$?", etc. As soon as we get a yes answer, we are done. On the the other hand, given that the previous k answers have all been NO, then the entropy of the next question, given the current state of knowledge, is precisely 1 (since with probability 1/2 X will be heads on the next flip). Thus, each question gives us 1 bit of additional information, which is the best we can do.

In the special case of fair coin flips, $E[\text{number of questions}] = E[X] = 2 = H(X)$. In general, $E[X]$ has nothing to do with $H(X)$. On the other hand, $E[\text{number of questions}]$ has a lot to do with $H(X)$. We will see later for *any* discrete random variable X, $H(X)$ represents the minimum number of questions required, on average, to ascertain the value of X.

(c) Intuitively, (X_1, X_2) has more information than $Y = X_1 + X_2$; hence $H(Y) < H(X_1, X_2)$.

Since $(X_1, X_2) \mapsto X_1 + X_2$ is not a one-to-one mapping, some states will get merged with a resultant loss of entropy. Hence, by the same argument as in Question 3, $H(X_1 + X_2) < H(X_1, X_2)$.

Alternatively, one can observe that by the chain rule,

$$\begin{aligned} H(X_1, Y) &= H(X_1) + H(Y|X_1) \\ &= H(X_1) + H(X_1 + X_2|X_1) \\ &= H(X_1) + H(X_2|X_1) \\ &= H(X_1, X_2) \end{aligned}$$

On the other hand, $H(X_1, Y) = H(Y) + H(X_1|Y) > H(Y)$ since Y does not completely determine X_1 and hence $H(X_1|Y) > 0$. Therefore, $H(Y) < H(X_1, X_2) = 2H(X)$, where the last equality follows from the fact that X_1 and X_2 are i.i.d., i.e. $H(X_1, X_2) = H(X_1) + H(X_2|X_1) = H(X_1) + H(X_2) = 2H(X_1)$.

6. **Two looks.**
Let X, Y_1, and Y_2 be binary random variables. Assume that $I(X; Y_1) = 0$ and $I(X; Y_2) = 0$.

(a) Does it follow that $I(X; Y_1, Y_2) = 0$? Prove or provide a counterexample.

(b) Does it follow that $I(Y_1; Y_2) = 0$? Prove or provide a counterexample.

Solution: Two looks

(a) The answer is “no”. Although at first the conjecture seems reasonable enough– after all, if Y_1 gives you no information about X, and if Y_2 gives you no information about X, then why should the two of them together give any information? But remember, it is NOT the case that $I(X; Y_1, Y_2) = I(X; Y_1) + I(X; Y_2)$. The chain rule for information says instead that $I(X; Y_1, Y_2) = I(X; Y_1) + I(X; Y_2|Y_1)$. The chain rule gives us reason to be skeptical about the conjecture.

This problem is reminiscent of the well-known fact in probability that pair-wise independence of three random variables is not sufficient to guarantee that all three are mutually independent. $I(X; Y_1) = 0$ is equivalent to saying that X and Y_1 are independent. Similarly for X and Y_2. But just because X is pairwise independent with each of Y_1 and Y_2, it does not follow that X is independent of the vector (Y_1, Y_2).

Here is a simple counterexample. Let Y_1 and Y_2 be independent fair coin flips. And let $X = Y_1$ XOR Y_2. X is pairwise independent of both Y_1 and Y_2, but obviously not independent of the vector (Y_1, Y_2), since X is uniquely determined once you know (Y_1, Y_2).

(b) Again the answer is "no". Y_1 and Y_2 can be arbitrarily dependent with each other and both still be independent of X. For example, let $Y_1 = Y_2$ be two observations of the same fair coin flip, and X an independent fair coin flip. Then $I(X; Y_1) = I(X; Y_2) = 0$ because X is independent of both Y_1 and Y_2. However, $I(Y_1; Y_2) = H(Y_1) - H(Y_1|Y_2) = H(Y_1) = 1$.

7. **Directed Information. [Bonus]**
Let $X^n = (X_1, \ldots, X_n)$ and $Y^n = (Y_1, \ldots, Y_n)$ be two sequences of random variables.

(a) Show that

$$I(X^n; Y^n) = \sum_{i=1}^{n} I(X^n; Y_i | Y^{i-1})$$

We define the directed information from X^n to Y^n by replacing the X^n in the sum by X^i, i.e.,

$$I(X^n \to Y^n) = \sum_{i=1}^{n} I(X^i; Y_i | Y^{i-1})$$

Directed information plays an important role in settings where causality plays a role, such as the capacity of channels with feedback. For now, we will show some properties of this quantity.

(b) Show that

$$I(X^n; Y^n) \geq I(X^n \to Y^n)$$

Hint: Use the data processing inequality shown in question 4(b).

(c) *Preservation Law*: Show that

$$I(X^n; Y^n) = I(X^n \to Y^n) + I(Y^{n-1} \to X^n)$$

where $Y^{n-1} = (0, Y_1, \ldots, Y_{n-1})$ and

$$I(Y^{n-1} \to X^n) = \sum_{i=1}^{n} I(Y^{i-1}; X_i | X^{i-1})$$

(d) Think about how you might interpret the definition of directed information and the results in parts (b) and (c).

Solution: Directed Information

(a) Straightforward extension of the property shown at the end of Lecture note 3 from last year (`https://web.stanford.edu/class/ee376a/files/2017-18/lecture_3.pdf`).

(b) Note that $Y_i - (Y^{i-1}, X^n) - (Y^{i-1}, X^i)$ forms a Markov triplet because given (Y^{i-1}, X^n), (Y^{i-1}, X^i) is deterministic and hence (Y^{i-1}, X^i) is conditionally independent of Y_i given (Y^{i-1}, X^n). Using result from question 4(b), we obtain

$$H(Y_i|Y^{i-1}, X^n) \leq H(Y_i|Y^{i-1}, X^i)$$

and hence

$$H(Y_i) - H(Y_i|Y^{i-1}, X^n) \geq H(Y_i) - H(Y_i|Y^{i-1}, X^i)$$

and

$$I(X^n; Y_i|Y^{i-1}) \geq I(X^i; Y_i|Y^{i-1})$$

Summing over i gives us the desired result.

(c) Starting by expanding the right hand side (RHS)

$$RHS = \sum_{i=1}^{n} \left\{ I(X^i; Y_i|Y^{i-1}) + I(Y^{i-1}; X_i|X^{i-1}) \right\}$$

Expanding the mutual information terms,

$$RHS = \sum_{i=1}^{n} \left\{ H(Y_i|Y^{i-1}) - H(Y_i|X^i, Y^{i-1}) + H(X_i|X^{i-1}) - H(X_i|Y^{i-1}, X^{i-1}) \right\}$$

Rearranging terms and splitting X^i into X^{i-1}, X_i,

$$RHS = \sum_{i=1}^{n} H(Y_i|Y^{i-1}) + \sum_{i=1}^{n} H(X_i|X^{i-1}) - \sum_{i=1}^{n} \left\{ H(Y_i|X_i, X^{i-1}, Y^{i-1}) + H(X_i|Y^{i-1}, X^{i-1}) \right\}$$

Using chain rule,

$$RHS = H(Y^n) + H(X^n) - \sum_{i=1}^{n} H(Y_i, X_i|X_i, X^{i-1}, Y^{i-1})$$

Chain rule again,

$$RHS = H(Y^n) + H(X^n) - H(X^n, X^n)$$

Using one of the equivalent definitions of mutual information,

$$RHS = I(X^n; Y^n)$$

This completes the proof.

(d) Directed information attempts to capture the "causal dependence" from X^n to Y^n. For example the term $I(X^i; Y_i|Y^{i-1})$ in the definition is the additional information about Y_i (given the past Y^{i-1}) by X^i (causally). Since this information can only be less than or equal to the additional information given by the entire X^n sequence, the result in part (b) is understandable.

To understand the result in part (c), consider a *channel* from X^n to Y^n that outputs Y_i given X_i according to some probabilistic law. Now suppose that there is no feedback, i.e., the transmitted inputs X_i do not directly depend on the previously received outputs Y^{i-1} given the past inputs X^{i-1}. In such a case, $I(Y^{i-1}; X_i|X^{i-1}) = 0$ which means $I(Y^{n-1} \to X^n) = 0$ and $I(X^n; Y^n) = I(X^n \to Y^n)$. Thus, in the case of no feedback the entire mutual information can be expressed in a causal manner. Thus the term $I(Y^{n-1} \to X^n)$ can be interpreted as the feedback information exploited by the transmitter and the term $I(X^n \to Y^n)$ can be thought of as the information flowing through the channel.

It should be noted that the real significance of quantities such as entropy, mutual information and directed information are due to their connection to real-life (operational) quantities such as optimal compression rate and channel capacity. But having an intuitive sense is also very helpful.

8. **Infinite entropy. [Bonus]**
This problem shows that the entropy of a discrete random variable can be infinite. (In this question you can take log as the natural logarithm for simplicity.)

(a) Let $A = \sum_{n=2}^{\infty}(n \log^2 n)^{-1}$. Show that A is finite by bounding the infinite sum by the integral of $(x \log^2 x)^{-1}$.

(b) Show that the integer-valued random variable X distributed as:
$P(X = n) = (An \log^2 n)^{-1}$ for $n = 2, 3, \ldots$ has entropy $H(X)$ given by:

$$H(X) = \log A + \sum_{n=2}^{\infty} \frac{1}{An \log n} + \sum_{n=2}^{\infty} \frac{2 \log \log n}{An \log^2 n}$$

(c) Show that the entropy $H(X) = +\infty$ (by showing that the sum $\sum_{n=2}^{\infty} \frac{1}{n \log n}$ diverges).

Solution: Infinite entropy.
We use the technique of bounding sums by integrals, see `https://math.stackexchange.com/questions/1282807/bounding-a-summation-by-an-integral` for an example with some figures.

(a) Define a function $f : [2, \infty) \to \mathbb{R}$ as follows:

$$f(x) = (\lceil x \rceil \log^2 \lceil x \rceil)^{-1}$$

Then, $f(x) \leq (x \log^2 x)^{-1}$ and

$$\begin{aligned} A &= (2\log^2 2)^{-1} + \sum_{n=3}^{\infty} (n \log^2 n)^{-1} \\ &= (2\log^2 2)^{-1} + \int_2^{\infty} (\lceil x \rceil \log^2 \lceil x \rceil)^{-1} dx \\ &\leq (2\log^2 2)^{-1} + \int_2^{\infty} (x \log^2 x)^{-1} dx \\ &= (2\log^2 2)^{-1} + \frac{1}{\log 2} \\ &< \infty \end{aligned}$$

(b) By definition, $p_n = \Pr(X = n) = 1/An\log^2 n$ for $n \geq 2$. Therefore

$$\begin{aligned} H(X) &= -\sum_{n=2}^{\infty} p_n \log p_n \\ &= -\sum_{n=2}^{\infty} \big(1/An\log^2 n\big) \log \big(1/An\log^2 n\big) \\ &= \sum_{n=2}^{\infty} \frac{\log(An\log^2 n)}{An\log^2 n} \\ &= \sum_{n=2}^{\infty} \frac{\log A + \log n + 2\log\log n}{An\log^2 n} \\ &= \log A + \sum_{n=2}^{\infty} \frac{1}{An\log n} + \sum_{n=2}^{\infty} \frac{2\log\log n}{An\log^2 n}\,. \end{aligned}$$

(c) The first term is finite. For base 2 logarithms, all the elements in the sum in the last term are nonnegative. (For any other base, the terms of the last sum eventually all become positive.) So all we have to do is bound the middle sum, which we do by comparing with an integral (in a similar manner as done in part (a), here using $\lfloor x \rfloor$ instead of $\lceil x \rceil$).

$$\sum_{n=2}^{\infty} \frac{1}{An\log n} = \int_2^{\infty} \frac{1}{A\lfloor x \rfloor \log \lfloor x \rfloor} dx > \int_2^{\infty} \frac{1}{Ax\log x} dx = K \ln\ln x \Big|_2^{\infty} = +\infty\,.$$

We conclude that $H(X) = +\infty$.