Homework Set #1
Properties of Entropy and Mutual Information

1. **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:

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

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 fixed, 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., $g(.)$ is one-to-one. Hence $H(X, g(X)) \geq H(g(X))$.

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

2. **Example of joint entropy.**
Let $p(x, y)$ be given by

X \ Y	0	1
0	$\frac{1}{3}$	$\frac{1}{3}$
1	0	$\frac{1}{3}$

Find

(a) $H(X), H(Y)$.
(b) $H(X|Y), H(Y|X)$.
(c) $H(X,Y)$.
(d) $H(Y) - H(Y|X)$.
(e) $I(X;Y)$.

3. **"True or False" questions**

Copy each relation and write **true** or **false**. Then, if it's true, prove it. If it is false give a counterexample or prove that the opposite is true.

(a) $H(X) \geq H(X|Y)$
(b) $H(X) + H(Y) \leq H(X,Y)$
(c) Let X, Y be two independent random variables. Then

$$H(X - Y) \geq H(X).$$

4. **Solution to "True or False" questions** e.

(a) $H(X) \geq H(X|Y)$ is **true**. Proof: In the class we showed that $I(X;Y) > 0$, hence $H(X) - H(X|Y) > 0$.
(b) $H(X) + H(Y) \leq H(X,Y)$ is **false**. Actually the opposite is true, i.e., $H(X) + H(Y) \geq H(X,Y)$ since $I(X;Y) = H(X) + H(Y) - H(X,Y) \geq 0$.
(c) Let X, Y be two independent random variables. Then

$$H(X - Y) \geq H(X).$$

True

$$H(X - Y) \overset{(a)}{\geq} H(X - Y|Y)) \overset{(b)}{\geq} H(X)$$

(a) follows from the fact that conditioning reduces entropy.
(b) Follows from the fact that given Y, $X - Y$ is a Bijective Function.

5. **Bytes.**
The entropy, $H_a(X) = -\sum p(x)\log_a p(x)$ is expressed in bits if the logarithm is to the base 2 and in bytes if the logarithm is to the base 256. What is the relationship of $H_2(X)$ to $H_{256}(X)$?

Solution: Bytes.

$$\begin{aligned}
H_2(X) &= -\sum p(x)\log_2 p(x) \\
&= -\sum p(x)\frac{\log_2 p(x)\log_{256}(2)}{log_{256}(2)} \\
&\overset{(a)}{=} -\sum p(x)\frac{\log_{256} p(x)}{log_{256}(2)} \\
&= \frac{-1}{\log_{256}(2)}\sum p(x)\log_{256} p(x) \\
&\overset{(b)}{=} \frac{H_{256}(X)}{\log_{256}(2)},
\end{aligned}$$

where (a) comes from the property of logarithms and (b) follows from the definition of $H_{256}(X)$. Hence we get

$$H_2(X) = 8H_{256}(X).$$

Solution: Example of joint entropy

(a) $H(X) = \frac{2}{3}\log\frac{3}{2} + \frac{1}{3}\log 3 = .918$ bits $= H(Y)$.
(b) $H(X|Y) = \frac{1}{3}H(X|Y=0) + \frac{2}{3}H(X|Y=1) = .667$ bits $= H(Y|X)$.
(c) $H(X,Y) = 3 \times \frac{1}{3}\log 3 = 1.585$ bits.
(d) $H(Y) - H(Y|X) = .251$ bits.
(e) $I(X;Y) = H(Y) - H(Y|X) = .251$ bits.

6. **Two looks.**
Here is a statement about pairwise independence and joint independence. Let X, Y_1, and Y_2 be binary random variables. If $I(X;Y_1) = 0$ and $I(X;Y_2) = 0$, does it follow that $I(X;Y_1,Y_2) = 0$?

(a) Yes or no?

(b) Prove or provide a counterexample.

(c) If $I(X;Y_1) = 0$ and $I(X;Y_2) = 0$ in the above problem, does it follow that $I(Y_1;Y_2) = 0$? In other words, if Y_1 is independent of X, and if Y_2 is independent of X, is it true that Y_1 and Y_2 are independent?

Solution: Two looks.

(a) The answer is "no".

(b) 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 $X = Y_1 \text{ XOR } Y_2$. 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).

(c) 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. **A measure of correlation.**
Let X_1 and X_2 be *identically distributed*, but not necessarily independent. Let

$$\rho = 1 - \frac{H(X_2|X_1)}{H(X_1)}.$$

(a) Show $\rho = \frac{I(X_1;X_2)}{H(X_1)}$.

(b) Show $0 \leq \rho \leq 1$.

(c) When is $\rho = 0$?

(d) When is $\rho = 1$?

Solution: A measure of correlation.

X_1 and X_2 are identically distributed and

$$\rho = 1 - \frac{H(X_2|X_1)}{H(X_1)}$$

(a)

$$\begin{aligned} \rho &= \frac{H(X_1) - H(X_2|X_1)}{H(X_1)} \\ &= \frac{H(X_2) - H(X_2|X_1)}{H(X_1)} \quad \text{(since } H(X_1) = H(X_2)\text{)} \\ &= \frac{I(X_1;X_2)}{H(X_1)}. \end{aligned}$$

(b) Since $0 \leq H(X_2|X_1) \leq H(X_2) = H(X_1)$, we have

$$0 \leq \frac{H(X_2|X_1)}{H(X_1)} \leq 1$$

$$0 \leq \rho \leq 1.$$

(c) $\rho = 0$ iff $I(X_1;X_2) = 0$ iff X_1 and X_2 are independent.

(d) $\rho = 1$ iff $H(X_2|X_1) = 0$ iff X_2 is a function of X_1. By symmetry, X_1 is a function of X_2, i.e., X_1 and X_2 have a one-to-one correspondence. For example, if $X_1 = X_2$ with probability 1 then $\rho = 1$. Similarly, if the distribution of X_i is symmetric then $X_1 = -X_2$ with probability 1 would also give $\rho = 1$.