TD 1

Exercise 1. In the Death Valley:

1. it rains on average once in 100 days;
2. the weather forecast foresees 3 rainy days in 100 days;
3. each time it rains, the weather forecast is right;
4. Guy-who-knows-everything predicts that it never rains.

Who is better, Guy-who-knows-everything or the weather forecast ?

Solution Intuitively, while the weather forecast is correlated with the weather that occurs, the weather predictions of the Guy-who-knows-everything are independent from the weather. We therefore expect that mutual information between the weather and the predictions of the Guy-who-knows-everything is equal to zero whereas there is non zero mutual information between the weather and the weather forecast. In this sense the weather forecast is better than the Guy-who-knows-everything.

More precisely, let X be the weather (0 if it is sunny, 1 when it rains),

Y be the weather forecast,

Z be the prediction of the Guy-who-knows-everything. The joint laws of X and Y and X and Z respectively are shown in the following tables

Table 1: The joint law of (X, Y), is given by reading the table as follows. At the "point" (a, b), $p(X = a, Y = b)$ is given.

$X \backslash Y$	0	1
0	0,97	0,02
1	0	0,01

Table 2: The joint law of (X, Z)

$X \backslash Z$	0	1
0	0,99	0
1	0,01	0

These tables were obtaining by reasoning as follows

Law of X: $p(X = 0) = 0,99$, $P(X = 1) = 0,01$

Law of Y: $p(Y = 0) = 0,97$, $P(Y = 1) = 0,03$

From the third point "each time it rains, the weather forecast is right" we deduce that $P(Y = 1|X = 1) = 1$ and $P(Y = 0|X = 1) = 0$. Therefore :

$$P(X = 1, Y = 1) = P(Y = 1|X = 1)P(X = 1) = 0,01$$

and

$$P(X = 1, Y = 0) = 0.$$

From this we obtain that

$$\begin{aligned} P(X=0,Y=1) &= P(Y=1) - P(X=1,Y=1) = 0,03 - 0,01 = 0,02 \\ P(X=0,Y=0) &= P(X=0) - P(X=0,Y=1) = 0,99 - 0,02 = 0,97 \end{aligned}$$

The second law is obtained in a similar way by observing that

$$P(X=0,Z=1) = P(X=1,Z=1) = 0.$$

Give $I(X;Y), I(X,Z)$: a simple computation leads to

$$\begin{aligned} I(X;Y) &\approx 0.05324 \\ I(X;Z) &= 0. \end{aligned}$$

This result has the following interpretation. The weather forecast is significantly correlated to the weather as shown by the ratio $\frac{I(X;Y)}{H(X)}$ that is a quantity between 0 and 1, a value of 0 indicating that X and Y are independent and a value of 1 indicating that X is a deterministic function of Y (see exercise 4). In our case

$$\frac{I(X;Y)}{H(X)} \approx 0.65902$$

whereas

$$\frac{I(X;Z)}{H(X)} = 0.$$

In the second case X and Z are independent and Z conveys no information about X.

Exercise 2. Let $(X_1, X_2, \ldots, X_N)$ be an N-tuple of random variables. Its entropy is defined as

$$H(X_1X_2\ldots X_N) = -\sum p(x_1,\ldots,x_N)\log p(x_1,\ldots,x_N)$$

1. (independent r.v.) Show that if $X_1,\ldots,X_n$ are independent, then

$$H(X_1X_2\ldots X_N) = \sum_{i=1}^{N} H(X_i)$$

2. (general case) Show the "chain rule for entropy"

$$H(X_1X_2\ldots X_N) = H(X_N|X_1,\ldots X_{N-1}) + H(X_{N-1}|X_1\ldots X_{N-2}) + \cdots + H(X_2|X_1) + H(X_1)$$

Solution

1. By induction. The formula is clearly true for $N=1$. Assume that the formula holds for $N=k$. Let $X=(X_1,\ldots,X_k)$ and $Y=X_{k+1}$. Observe that X and Y are independent. We obtain

$$\begin{aligned} H(X_1,\ldots,X_{k+1}) &= H(X,Y) \\ &= H(X) + H(Y) - I(X;Y) \\ &= H(X) + H(Y) \text{ (because } X \text{ and } Y \text{ are independent)} \\ &= H(X_1,\ldots,X_k) + H(X_{k+1}) \\ &= H(X_1) + \cdots + H(X_k) + H(X_{k+1}) \text{ (by induction)} \end{aligned}$$

2. By induction again. The formula clearly holds for $N=1$. Let $X=(X_1,\ldots,X_k)$ and $Y=X_{k+1}$.

$$\begin{aligned} H(X_1,\ldots,X_{k+1}) &= H(X,Y) \\ &= H(X) + H(Y|X) \\ &= H(X_1,\ldots,X_k) + H(X_{k+1}|X_1\ldots X_k) \\ &= H(X_1) + \cdots + H(X_k|X_1\ldots X_{k-1}) + H(X_{k+1}|X_1\ldots X_k) \text{ (by induction)} \end{aligned}$$

Exercise 3. Let (X, Y, Z) be a tuple of random variables such that

$$\begin{aligned} p_{XYZ}(0,0,0) &= \frac{1}{4} \\ p_{XYZ}(0,1,0) &= \frac{1}{4} \\ p_{XYZ}(1,0,0) &= \frac{1}{4} \\ p_{XYZ}(1,0,1) &= \frac{1}{4} \end{aligned}$$

Recall that

$$H(Y|X) = \sum_x p(x) H(Y|X = x)$$

Compute $H(X)$, $H(Y|X)$, $H(Z|X,Y)$. Find $H(X,Y,Z)$ in two different manners (chain rule and direct computation). Recall that $h(1/4) \approx 0.811$. Compute $H(Y)$ and verify that $H(Y|X) \leq H(Y)$. How much information conveys X about Y and how much information conveys Y about X ?

Solution

- We have $p(X = 0) = p(X = 1) = 1/2$. Therefore $H(X) = 1$.
- Ensuite $p(Y = 0|X = 0) = 1/2$ et $p(Y = 1|X = 0) = 1/2$. Donc $H(Y|X = 0) = 1$. De plus $p(Y = 1|X = 1) = 0$, et $p(Y = 0|X = 1) = 1$, donc $H(Y|X = 1) = 0$. Finally

$$H(Y|X) = 1/2 H(Y|X = 0) + 1/2 H(Y|X = 1) = 1/2.$$

- $p(Z = 0|X = 0, Y = 0) = 1$, $p(Z = 1|X = 0, Y = 0) = 0$. Therefore $H(Z|X = 0, Y = 0) = 0$. By a similar computation we see that $p(Z = 0|X = 0, Y = 1) = 1$ $p(Z = 1|X = 0, Y = 1) = 0$. Therefore $H(Z|X = 0, Y = 1) = 0$. Finally $H(Z|X = 1, Y = 0) = 1$. From this we deduce that

$$H(Z|X,Y) = 1/4.0 + 1/4.0 + 1/2 H(Z|X = 1, Y = 0) = 1/2.$$

- $H(X,Y,Z) = H(Z|X,Y) + H(Y|X) + H(X) = 1/2 + 1/2 + 1 = 2$.
- Direct computation : the random variable (X,Y,Z) is uniformly distributed over 4 values, therefore $H(X,Y,Z) = \log_2 4 = 2$
- Observe that $p(Y = 0) = \frac{3}{4}$ and $p(Y = 1) = \frac{1}{4}$. It follows that $H(Y) = h(1/4) \approx 0.811$. $H(Y|X) = \frac{1}{2}$. This is much smaller than $H(Y)$. The information brought by X on Y is equal to $I(X;Y) = H(Y) - H(Y|X) \approx 0.311$. The information brought by Y on X is also equal to $I(X;Y)$.

Exercise 4. Show that $H(Y|X) = 0$ if and only if Y is a function of X, that is for all x, such that $p(x) > 0$ there exists y such that $p(y|x) = 1$. We use here the notation $p(y|x) = p(Y = y|X = x)$.

Solution We have $H(Y|X) = \sum_x p(x) H(Y|X = x)$. This implies that for all x such that $p(x) > 0$, $H(Y|X = x) = 0$. This implies that $p(y|x) = 1$ for a certain y and $p(y|x) = 0$ for the others.

Exercise 5. Let X be a random variable and $g(x)$ be a function. By using the chain rule in two different ways, show that

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

Solution We write

$$\begin{aligned} H(X, g(X)) &= H(g(X)|X) + H(X) \\ &= H(X) \end{aligned}$$

On the other hand

$$\begin{aligned} H(X, g(X)) &= H(X|g(X)) + H(g(X)) \\ &\geq H(g(X)) \end{aligned}$$

Exercise 6. Let X be a random variable with entropy $H(X)$. What is the relationship between $H(Y)$ and $H(X)$

- when $Y = 2^X$
- when $Y = \cos X$.

Solution

1. Y is a function of X hence $H(Y) \leq H(X)$ from the previous exercise. Moreover X is a function of Y, because $X = \log_2 Y$. Therefore $H(X) \leq H(Y)$. All this implies that $H(X) = H(Y)$. Thsi proof technique can be used to show that $H(X) = H(f(X))$ when f is bijective.

2. For the same reason $H(Y) \leq H(X)$. However in this case the cosine function is not bijective, and it is not always true that $H(X) = H(Y)$. For instance for a random variable X taking its values in $\{-1, 1\}$ with $p(X = -1) = p(X = 1) = \frac{1}{2}$, we have $H(X) = 1$. However here $p(Y = \cos 1) = 1$, which implies that $H(Y) = 0$.

Exercise 7. We assume here that we have a twin-pan scale (which only allows to compare two objects that have been put on the pans) and n coins. A genuine coin has a certain weight w which is unknown. A counterfeit coin has a different weight, however this weight can be either smaller or larger than the weight of a genuine coin. We know that that among these n coins there is at most one counterfeit coin.

1. Give the best lower bound you can find on the number of weighings that have to be performed to detect with certainty a counterfeit coin. *Hint: use Exercise 5...*

2. Assume now that we also have an auxiliary infinite stack of genuine coins. Give the optimal strategy to find the counterfeit coin if there is one when $n = 13$. Generalize your strategy to $n = \frac{3^k-1}{2}$.

Solution

1. Number the coins from 1 to n. We encode every possibility by a random variable X which takes integer values from $-n$ to n. X is equal 0 when there is no counterfeit coin and otherwise $|X|$ is equal to the number of the counterfeit coin. A positive value for X means that the counterfeit is heavier, where a negative values means that the counterfeit coin is lighter. For instance $X = 2$ means that the second coin is the counterfeit coin and that this coin is heavier than genuine coins.

   Let k be the number of weighings after which we know with certainty X. Let $(Y_1, Y_2, \ldots, Y_k) \in \{-1, 0, 1\}^k$ be the result of these k weighings. $Y_i = 1$ means that the i-th weighing gave the answer that what put on the left pan of the scale was heavier than what was put on the right pan. $X_i = -1$ means the opposite: the right part was heavier, whereas $X_i = 0$ means that both parts have the same weight.

By assumption X is a function of $(Y_1, Y_2, \ldots, Y_k)$. Therefore (exercise 5)

$$H(X) \leq H(Y_1, \ldots, Y_k).$$

Moreover

$$\begin{aligned} H(Y_1, \ldots, Y_k) &= H(Y_1) + H(Y_2|Y_1) + \cdots + H(Y_k|Y_1 \ldots Y_{k-1}) \text{ (chaining)} \\ &\leq H(Y_1) + H(Y_2) + \cdots + H(Y_k) \text{ (conditioning can only reduce entropy)} \\ &\leq k \log_2 3 \text{ ($\log_2 3$ is the maximum entropy of a ternary variable)} \end{aligned}$$

Hence

$$k \geq \frac{H(X)}{\log_2 3}.$$

This lower bound is maximized for an appropriate choice of X- in this case X should be uniformly distributed to maximize entropy. In such a case

$$H(X) = \log_2(2n+1).$$

We eventually obtain

$$k \geq \frac{\log_2(2n+1)}{\log_2 3} = \log_3(2n+1).$$

2. The previous bound can also be applied with an auxiliary heap of genuine coins. The lower bound tells us that we need at least 3 weighings. Consider as a starter the case of 1 coin. In this case it suffices to check whether this coin is heavier or not than a genuine coin.

   Another case that is interesting is the case of 4 coins. In such a case the lower bound tells us that we need at least 2 weighings. This can be achieved as follows. Take 3 coins that are compare to 3 genuine coins. There are three cases to consider :
   (a) both sets have the same weight then it means that all three coins are genuine and it suffices to check whether the last coin is genuine in one weighing by comparing it to a genuine coin.
   (b) the three coins are heavier than 3 genuine coins. Then we know two things: (i) this set of three coins contains a counterfeit coin, (ii) a counterfeit coin is heavier than a genuine coin.
   This last point simplifies the problem since we just have to compare the weight of two of these 3 coins. If they have the same weight, then it means that the remaining coin is the counterfeit coin, otherwise the heaviest of these 2 coins is the counterfeit coin.
   (c) the three coins are lighter than 3 genuine coins. The answer to this case is basically the same as the previous one (by replacing heavier by lighter).

   This strategy can be generalized to 13 coins by taking 9 coins from this set and comparing them to 9 genuine coins. If they have the same weight as 9 genuine coins, then it means that the counterfeit coin (if there is one) is in remaining set of 4 coins. This can be checked in just 2 weighings by the previous strategy. If the 9 coins are heavier, then this means again two things (i) this set of 9 coins contains a counterfeit coin, (ii) a counterfeit coin is heavier than a genuine coin.

   The counterfeit coin can be found in two weighings by taking 6 coins from these 9 coins, divide them in two sets of 3 coins and compare their weight. If they have the same weight this means that the counterfeit coin is in the remaining 3 coins and this can be checked in just one weighing as shown previously. Otherwise the counterfeit coin is among the set of 3 coins which is the heaviest and the counterfeit coin can be found with just one additional weighing.

   This algorithm can be generalized to every ensemble of $\frac{3^k-1}{2}$ coins. This needs k weighings. This can be shown by induction on k by using the auxiliary result that a heavier counterfeit

coin hidden in a set of 3^k coins (the rest is formed by genuine coins can be found in just k weighings (by generalizing the previous approach and dividing the set in 3 at each step). This can be done as follows. We choose 3^{k-1} coins. We compare them to 3^{k-1} genuine coins and if the two sets have a different weight, then we know that this set of 3^{k-1} contains a counterfeit coin and whether this counterfeit coin is heavier or lighter. We use the auxiliary result to find this counterfeit coin in just $k-1$ additional weighings. On the other hand if the two sets have the same weight, then this means that the counterfeit coin if it exists can only be amont the remaining $\frac{3^k-1}{2} - 3^{k-1} = \frac{3^{k-1}-1}{2}$ coins. We use induction to check this remaining set in just $k-1$ weighings.

Exercise 8 (The password problem)**.** A hacker wants to access a server which is protected by a password that is unknown to the hacker. We assume that the passwords are binary sequences of length m and we let $\mathcal{M} = \{0,1\}^m$ be the set of all possible passwords. We also assume that the only way for the hacker to access the server is find the right password by checking them one by one.

It turns out that the password is chosen randomly in $\mathcal{M}$ according to a probability distribution of entropy $h \leq m$. We denote by p_i the probability of the i-th element of $\mathcal{M}$ and we assume that the order on $\mathcal{M}$ is chosen in such a way that

$$p_1 \geq p_2 \geq p_3 \cdots \geq p_{2^m}.$$

1. Show that the best strategy consists in trying the elements according to the aforementioned order, i.e. try first the element of index 1, that is the element of probability p_1, then the second element and so on and so forth. Give a formula for $\mathcal{N}(p)$ which is the expected number of passwords that have to be tested before finding the right one.

2. Consider now two probability distributions over the positive integers $p = (p_i)_{i\geq 1}$ and $q = (q_i)_{i\geq 1}$ such that the p_i's and the q_i's are decreasing and $q_i > 0$ for all i. On the other hand p_i is allowed to be equal to zero for i large enough. Recall that the Kullback distance from p to q is given by
$$D(p \parallel q) = \sum_{i\geq 1} p_i \log \frac{p_i}{q_i}. \tag{1}$$
Let $q_i = (1-\alpha)\alpha^{i-1}$ for a certain real $0 < \alpha < 1$. Show that if $H(p) = H(q)$, then $\sum_{i\geq 1} ip_i \geq \sum_{i\geq 1} iq_i$.
Indication: use that $D(p \parallel q) \geq 0$.

3. Express $H(q)$ in terms of α. Let H_α be this expression. We recall that $\sum_{i\geq 1} \alpha^{i-1} = 1/(1-\alpha)$ and $\sum_{i\geq 1} i\alpha^{i-1} = 1/(1-\alpha)^2$

4. Deduce that for all $0 < \alpha < 1$ we have $1 < (1-\alpha)2^{H_\alpha} < e$, where e is the base of the natural logarithm.

5. Deduce from the previous result that $\mathcal{N}(p) > c_1 2^h$ (give c_1 explicitly here). Give an interpretation of this result.

Solution

1. We check all the passwords according to a certain order. More precisely, we choose a permutation π of the integers from 1 to 2^m, and we first check the password of index $\pi(1)$, then the password of index $\pi(2)$ and so on and so forth. The expected number of tests $N(\pi)$ is equal to
$$N(\pi) = \sum_{i=1}^{2^m} ip_{\pi(i)}$$

Assume that there exists a pair (i, j) such that $i < j$ and $\pi(i) > \pi(j)$. We can obtain a smaller number of tests by choosing another order (i.e. a permutation π') such that $\pi'(k) = \pi(k)$ for all k different from i and j and $\pi'(i) = \pi(j)$, $\pi'(j) = \pi(i)$. We observe namely that

$$\begin{aligned} N(\pi') - N(\pi) &= i(p_{\pi(j)} - p_{\pi(i)}) + j(p_{\pi(i)} - p_{\pi(i)}) \\ &= (j-i)(p_{\pi(i)} - p_{\pi(j)}) \\ &< 0. \end{aligned}$$

This implies that the trivial order $\pi(1) = 1, \pi(2) = 2, \ldots, \pi(2^m) = 2^m$ minimizes $N(\pi)$.

2. We first observe that

$$\log q_i = \log(1-\alpha) + (i-1)\log\alpha$$

Therefore :

$$\begin{aligned} i &= \frac{\log q_i - \log(1-\alpha) + \log\alpha}{\log\alpha} \\ &= \frac{\log q_i + \log\frac{\alpha}{1-\alpha}}{\log\alpha} \end{aligned}$$

We use this expression of i in the following computation

$$\begin{aligned} \sum_i i p_i &= \sum_i p_i \frac{\log q_i + \log\frac{\alpha}{1-\alpha}}{\log\alpha} \\ &= \frac{1}{\log\alpha}\sum p_i \log q_i + \frac{\log\frac{\alpha}{1-\alpha}}{\log\alpha} \\ &= \frac{1}{\log\alpha}\sum p_i \log\frac{q_i}{p_i} p_i + \frac{\log\frac{\alpha}{1-\alpha}}{\log\alpha} \\ &= -\frac{1}{\log\alpha} D(p \parallel q) - \frac{1}{\alpha} H(p) + \frac{\log\frac{\alpha}{1-\alpha}}{\log\alpha} \\ &\leq -\frac{1}{\log\alpha} H(q) + \frac{\log\frac{\alpha}{1-\alpha}}{\log\alpha} \\ &= \frac{\log\frac{\alpha}{1-\alpha}}{\log\alpha} + \frac{1}{\log\alpha}\sum_i q_i \log q_i \\ &= \frac{\log\frac{\alpha}{1-\alpha}}{\log\alpha} + \frac{1}{\log\alpha}\sum_i q_i \left(\log(1-\alpha) + (i-1)\log\alpha\right) \\ &= \sum_i i q_i \end{aligned}$$

3. The computation of the entropy of q gives

$$\begin{aligned} H_\alpha &= -\sum_i q_i \log q_i \\ &= -\sum_i (1-\alpha)\alpha^{i-1}\left(\log(1-\alpha) + (i-1)\log\alpha\right) \\ &= -\log(1-\alpha) + \log\alpha - \frac{\log\alpha}{1-\alpha} \\ &= \log\frac{\alpha}{1-\alpha} - \frac{\log\alpha}{1-\alpha} \end{aligned}$$

4. We have

$$\begin{aligned}(1-\alpha)2^{H_\alpha} &= (1-\alpha)\frac{\alpha}{1-\alpha}2^{-\frac{\log\alpha}{1-\alpha}} \\ &= \alpha\alpha^{-\frac{1}{1-\alpha}} \\ &= \alpha^{-\frac{\alpha}{1-\alpha}}\end{aligned}$$

It is straightforward to check that this expression is always between 1 and e.

5. We also observe that

$$\sum_i iq_i = \frac{1}{1-\alpha}.$$

From this we deduce that

$$\mathcal{N}(p) \geq \mathcal{N}(q) = \frac{1}{1-\alpha}.$$

where α is such that $h = H_\alpha$. From the inequality proved in the previous point we deduce that

$$2^{-H_\alpha} \leq 1-\alpha \leq e2^{-H_\alpha}.$$

Hence :

$$\frac{1}{e}2^{H_\alpha} \leq \frac{1}{1-\alpha} \leq 2^{H_\alpha},$$

that is

$$\frac{1}{e}2^{h} \leq \frac{1}{1-\alpha} \leq 2^{h}.$$

This implies that

$$\mathcal{N}(p) \geq \frac{1}{e}2^h.$$

Exercise 9. Show that

$$\binom{n}{t} \leq 2^{nh(t/n)}$$

Solution Let $(X_1, X_2, \ldots, X_n)$ be a random variable which is uniformly distributed among the binary words of length n and weight t. We have

$$H(X_1, X_2, \ldots, X_n) = \log_2 \binom{n}{t} \tag{2}$$

On the other hand,

$$\begin{aligned}H(X_1, X_2, \ldots, X_n) &\leq H(X_1) + H(X_2|X_1) + \cdots + H(X_n|X_1 \ldots Y_{n-1}) \text{ (chaining)} \\ &\leq H(X_1) + H(X_2) + \cdots + H(X_n) \text{ (conditioning can only reduce entropy)} \\ &\leq nh(t/n) \text{ (since } p(X_i = 1) = t/n \text{ and therefore } H(X_i) = h(t/n).)\end{aligned}$$

Exercise 10. [Fano's lemma- relationship between error probability of an estimator and conditional probability] Let X and Y be two random variables (with X taking its values in an alphabet of size a). Let $\hat{X}$ be an estimator for X computed from the knowledge of Y (it is therefore a function of Y). Denote by P_e the error probability of the estimator, that is $P_e = p(\hat{X} \neq X)$. Show that

$$h(P_e) + P_e \log_2(a-1) \geq H(X|Y).$$

Hint : introduce a random variable E defined by

$$E = \begin{cases} 1 & \text{si } \hat{X} \neq X \\ 0 & \text{si } \hat{X} = X \end{cases}$$

and find two different expressions for $H(E, X|Y)$. Prove that Fano's inequality is sharp by considering the following example. X takes a different values $1, \ldots, a$ with probability $p_i \stackrel{\text{def}}{=} \mathbf{Prob}(X = i)$ and

$$p_1 \geq p_2 = \cdots = p_a$$

and Y is independent of X.

Solution We write the entropy $H(E, X|Y)$ in two different ways

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

On one hand $H(E|X, Y) = 0$ (since E is a function of X and Y). On the other hand, concerning the second expression for $H(E, X|Y)$ we have $H(E|Y) \leq H(E)$ (conditioning can only reduce entropy). Since E is a binary random variable

$$H(E) = -p(E = 0) \log(p(E = 0)) - p(E = 1) \log(p(E = 1)) = h(P_e).$$

We also observe that

$$\begin{aligned} H(X|E, Y) &= p(E = 0)H(X|Y, E = 0) + p(E = 1)H(X|Y, E = 1) \\ &\leq (1 - P_e)0 + P_e \log_2(a - 1) \end{aligned}$$

because when $E = 0$, X is a function of Y and therefore $H(X|Y, E = 0) = 0$ and we can always bound the conditional entropy of X given Y and $E = 1$ by the entropy uniformly distributed among all possible values for X with the exception of $\hat{X}$. By gathering all these observations we obtain Fano's lemma.

For the case when Fano's lemma is sharp we obtain for a random variable X that takes a different values $1, \ldots, a$ with probability $p_i \stackrel{\text{def}}{=} \mathbf{Prob}(X = i)$,

$$p_1 \geq p_2 = \cdots = p_a$$

and Y which is independent of X the following conditional entropy

$$\begin{aligned} H(X|Y) &= H(X) \text{ (ind. of } X \text{ and } Y) \\ &= -p_1 \log p_1 - \sum_{i=2}^{a} p_i \log p_i \\ &= -p_1 \log p_1 - \sum_{i=2}^{a} \frac{1 - p_1}{a - 1} \log \left(\frac{1 - p_1}{a - 1} \right) \\ &= -p_1 \log p_1 - (a - 1) \frac{1 - p_1}{a - 1} \log \left(\frac{1 - p_1}{a - 1} \right) \\ &= -p_1 \log p_1 - (1 - p_1) \log \left(\frac{1 - p_1}{a - 1} \right) \\ &= -p_1 \log p_1 - (1 - p_1) \log(1 - p_1) + (1 - p_1) \log(a - 1) \\ &= h(P_e) + P_e \log(a - 1) \end{aligned}$$

Exercise 11. Let X and Y be two random variables taking their values in a group $(G, +)$. Let $Z = X + Y$.

1. Show that $H(Z|X) = H(Y|X)$.
2. Show that if X et Y are independent $H(Y) \leq H(Z)$ and $H(X) \leq H(Z)$ (use non-negativity of mutual information).
3. Give an example of two random variables X and Y such that $H(X) > H(Z)$ and $H(Y) > H(Z)$.

Solution

1. We write in two different ways $H(Z, X, Y)$:

$$\begin{aligned} H(Z,X,Y) &= H(Z|X,Y) + H(Y|X) + H(X) \\ &= 0 + H(Y|X) + H(X) \\ &= H(Y|X,Z) + H(Z|X) + H(X) \\ &= 0 + H(Z|X) + H(X) \end{aligned}$$

2. We observe that

$$\begin{aligned} H(Y) &= H(Y|X) (\text{because } X \text{ and } Y \text{ are independent}) \\ &= H(Z|X) \\ &\leq H(Z) \end{aligned}$$

3. Consider a non-deterministic binary random variable $X \in \{0, 1\}$ with arbitrary probabilities (meaning that $\mathbf{Prob}(X = 0) \neq 0$ and $\mathbf{Prob}(X = 1) \neq 0$) and $Y = 1 - X$. Then $Z = 0$ and $H(Z) = 0$.