
Solution of exercise sheet n° 2 :

2-1.

(a) First, one needs to find the marginal probability distributions $p(x)$ and $p(y)$. For this we use the relation $p(x) = \sum_y p(x,y)$. We obtain that $p(x) = p(y) = \{\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\}$. Thus, we find $H(X) = -\sum_x p(x) \log p(x) = H(Y) = \log 3$ bits.

(b) In order to compute $H(X|Y)$, we need to find $p(x|y)$, which we obtain via $p(x|y) = p(x,y)/p(y)$. Using the definition of $H(X|Y)$, we obtain $H(X|Y) = -\sum_{x,y} p(x,y) \log_2 p(x|y) = \log 3 - 4/9$ bits. With the same method, we find $H(Y|X) = \log 3 - 4/9$ bits.

(c) Using the results of (a) and (b), we directly compute $H(X,Y) = H(X) + H(Y|X) = 2\log 3 - 4/9$ bits.

(d) Using (a) and (b), we find $H(X{:}Y) = H(X) - H(X|Y) = 4/9$ bits.

(e) See lecture.

2-2.

(a) With the chain rule one one can write
$H(X_1, X_2, ..., X_k) = \sum_{i=1}^{k} H(X_i|X_{i-1}, ..., X_1)$.
The i-th draw with replacement implies that X_i is independent of X_j. Thus, $H(X_1, X_2, ..., X_k) = \sum_{i=1}^{k} H(X_i)$.
As all draws have the same probability distribution, we obtain
$H(X_1, X_2, ..., X_k) = kH(X)$.

(b) The random variable X_i corresponds to drawing a color at the i-th draw. Here, the calculation of the entropy $H(X_i)$ is not depenending on the previous draws, as we *did not obtain any information*. All balls up to the i-th draw were drawn *without looking at them*, that is, *without gain of information*. Therefore, the probabilities did not change. (The experiment can be described as taking $i-1$ balls from one urn and putting them into another urn without looking at them. Logically the probability distribution of the ith draw is not affected by that.)

(c) We find that $p(X_1 = c_1, X_2 = c_2) = p(X_1 = c_2, X_2 = c_1)$, where c_i is a certain color. Thus, a draw $c_1 c_2$ has the same probability as a draw $c_2 c_1$. A constructive way to understand this : Assume the total number of balls is $M = R + W + B$, where $R = p_R\, M$ is the number of red balls, $W = p_W\, M$ is the number of white balls, and $B = p_B\, M$ is the number of black balls. With this information it is possible to construct a tree, where on each branch one can write the probability to draw a certain color. Example : First draw is a red ball, probability is $p_R = R/M$. Second ball is a white ball, probability is $W/(M-1)$. Compare with firstly drawing a white ball (probability $p_W = W/M$), and secondly drawing a red ball (probability $R/(M-1)$). We find

$$\frac{R}{M}\frac{W}{M-1} = \frac{W}{M}\frac{R}{M-1}.$$

This example could be repeated for all color combinations, proving the relation.

(d) The probability to draw a red ball with the second draw is given by

$$p(X_2 = r) = p(X_1 = r, X_2 = r) + p(X_1 = b, X_2 = r) + p(X_1 = n, X_2 = r),$$

since getting a red ball for the second draw may be preceded by drawing a red, black or white ball first. With the result of (c) one finds

$$p(X_2 = r) = p(X_1 = r, X_2 = r) + p(X_1 = r, X_2 = b) + p(X_1 = r, X_2 = n) = p(X_1 = r).$$

(e) The calculation of (d) shows that $p(X_2 = r) = p(X_1 = r)$. Similarly, $p(X_2 = b) = p(X_1 = b)$ and $p(X_2 = n) = p(X_1 = n)$.

(f) The marginal probabilities are the same for the first or second draw, i.e. $p(X_2 = c_i) = p(X_1 = c_i)$, thus $H(X_2) = H(X_1)$.

(g) By using the chain rule $H(X_i|X_{i-1}, ..., X_1) \leq H(X_i)$, one can write for dependent random variables $H(X_1, X_2, ..., X_k) \leq \sum_{i=1}^{k} H(X_i)$.
Using $H(X_i) = H(X)$, one obtains $H(X_1, X_2, ..., X_k) \leq kH(X)$.

2-3.

(a) Using the definition of the conditional probability, one can write $p(x, z|y) = p(x|y)p(z|x, y)$. However, for the Markov chain $p(z|x, y) = p(z|y)$, thus one obtains $p(x, z|y) = p(x|y)p(z|y)$.

(b) The chain rule for mutual entropies is given by

$$H(X_1, X_2, ..., X_n{:}Y) = \sum_{i=1}^{n} H(X_i{:}Y|X_1, X_2, ..., X_{i-1}).$$

Thus, $H(X{:}Y, Z) = H(Y, Z{:}X) = H(Y{:}X) + H(Z{:}X|Y)$ and $H(Y, Z{:}X) = H(Z{:}X) + H(Y{:}X|Z)$. Furthermore, we have the definition (see lecture)

$$H(Z{:}X|Y) = -\sum_{xyz} p(x, y, z) \log \frac{p(x|y)p(z|y)}{p(z, x|y)}.$$

Using the result of (a), we conclude that $H(Z{:}X|Y) = 0$. Taking into account that $H(Y{:}X|Z) \geq 0$, one obtains $H(X{:}Y) \geq H(X{:}Z)$.

(c) Using the result of (b), $H(X{:}Z) \leq H(X{:}Y) = H(Y) - H(Y|X)$. Now $\max\{H(X{:}Y)\} = \log k$ as $H(Y|X) \geq 0$ and $\max\{H(Y)\} = \log k$. The limit is reached if $Y = f(X)$ and Y is uniformly distributed. One finally obtains the inequality $H(X{:}Z) \leq \log k$.

(d) If $k = 1$, then $H(X{:}Z) = 0$. The set $\mathcal{Y}$ contains only one element, thus all information contained in X is lost by the operation $X \to Y$.

2-4.

(a) The probability of a Bernoulli experiment in general reads $p(x_1, x_2, ...x_n) = p^k(1 - p)^{n-k}$. Since for a typical sequence $k \approx np$, we find the probability to emit a particular typical sequence : $p(x_1, x_2, ...x_n) = p^k(1 - p)^{n-k} \approx p^{np}(1 - p)^{n(1-p)}$.
We can approximate as a function of the entropy :

$$\log p(x_1, x_2, ...x_n) \approx np \log p + n(1 - p) \log(1 - p) = -nH(p).$$

Thus, $p(x_1, x_2, ...x_n) \approx 2^{-nH(p)}$.

(b) The number of typical sequences N_{ST} is given by the number of ways to have np ones in a sequence of length n (or to get np successes for n trials in a Bernoulli experiment). Thus,

$N_{ST} = \binom{n}{np} = \frac{n!}{(np)!(n(1-p))!}$.
By using the Stirling approximation one obtains $\log N_{ST} \approx nH(p)$.
Comparison to the total number of sequences that can be emitted by the source : $N_{ST} = 2^{nH(p)} \leq 2^n$.
The probability that the source emits a sequence that is typical is $P_{ST} = p_{ST}N_{ST} \approx 1$ for $n \gg 1$.

(c) The most probable sequence 1111.....1 if $p > 1/2$ or 0000.....0 if $p < 1/2$. This sequence is not typical.

2-5.

(a) By replacing $H(Y|X) = H(X,Y) - H(X)$ in the definition of the distance, we obtain $2H(X,Y) - H(X) - H(Y)$. Furthermore, the definition $H(X{:}Y) = H(X) + H(Y) - H(X,Y)$ gives us another expression for the distance.

(b) Proof of the properties in order of appearance :

1. $\rho(x,y) \geq 0$ since $H(X|Y) \geq 0$ and $H(Y|X) \geq 0$.
2. $\rho(x,y) = \rho(y,x)$ is trivially given by its definition.
3. $\rho(x,y) = 0$ iff $H(Y|X) = H(X|Y) = 0$, which holds iff there exists a bijection between X and Y.
4. Let $A = \rho(x,y)+\rho(y,z)-\rho(x,z)$. Using a), $A = 2[H(X,Y)+H(Y,Z)-H(Y)-H(X,Z)]$. Using the strong subadditivity $H(X,Y) + H(Y,Z) - H(Y) \geq H(X,Y,Z))$, we have $A \geq 2[H(X,Y,Z) - H(X,Z)] \equiv 2H(Y|X,Z) \geq 0$.

2-6.

(a) For instance if $\mathcal{X} = \mathcal{Y} = \mathcal{Z} = \{0,1\}$, $X = Y = Z$ with uniform distributions.
We have $H(X{:}Y) = 1$ bit since $H(X{:}Y) = H(Y) - H(Y|X)$ and $H(Y|X) = 0$ (because X are Y perfectly correlated). We find $H(X{:}Y|Z) = 0$ bit since $(X,Y) = f(Z)$. One verifies that $H(X{:}Y{:}Z) > 0$ and $H(X{:}Y|Z) < H(X{:}Y)$.

(b) For instance if $\mathcal{X} = \mathcal{Y} = \mathcal{Z} = \{0,1\}$ and $Z = X \oplus Y$ (sum mod 2), with :

P(X,Y)	$Y=$ 0	1	
0	1/4	1/4	1/2
$X=$ 1	1/4	1/4	1/2
	1/2	1/2	1

We obtain $H(X{:}Y) = 0$ bit since X and Y are independent and thus $H(Y|X) = H(Y)$. Furthermore, $H(X{:}Y|Z) = H(X|Z) - H(X|Y,Z)$. In our example X is fixed if one knows Y and Z. Thus, $H(X|Y,Z) = 0$. This implies $H(X{:}Y|Z) = H(X|Z)$. One obtains $H(X{:}Y|Z) = 1$ bit. One verifies that $H(X{:}Y{:}Z) = -1$ bit < 0 bit and $H(X{:}Y|Z) > H(X{:}Y)$. We confirm furthermore, that $H(X : Z) = H(Y : Z) = 0$. Therefore, the corresponding Venn diagram is like in Fig. 1, which shows that there is a *negative* overlap between the three random variables X, Y and Z.

Optional : An interesting exercise is to determine under which conditions (independence, perfect correlation) on the three variables X, Y and Z one obtains a maximal or minimal $H(X{:}Y{:}Z)$.

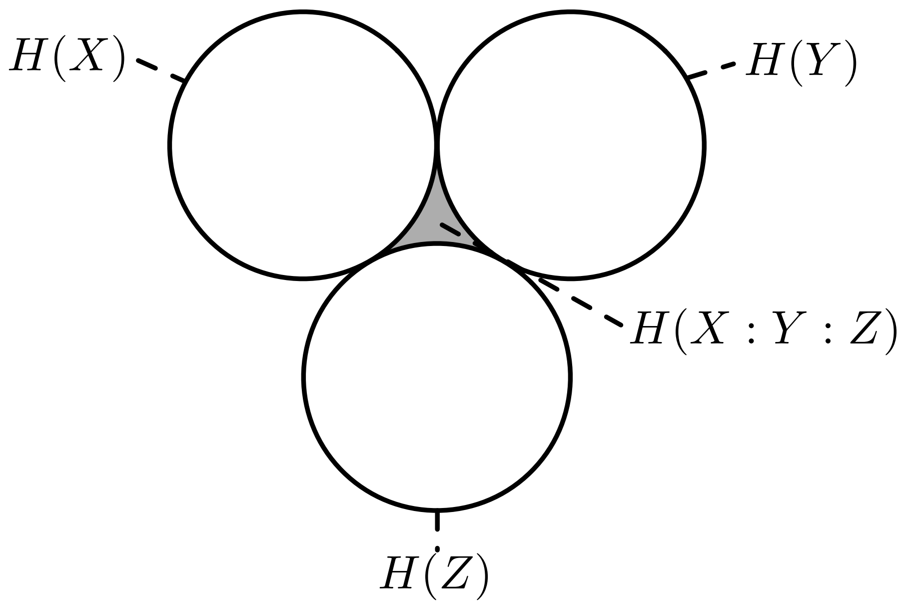

FIGURE 1 – Venn diagram depicting example 2-6. (b).