Solutions to Exercise Sheet 2

Exercise 1.

(a) First, we need to find the marginal probability distributions $p(x)$ and $p(y)$.
For this we use the relation $p(x) = \sum_y p(x, y)$, which gives $p(x) = p(y) = \{\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\}$.
Therefore $H(X) = -\sum_x p(x) \log p(x) = H(Y) = \log 3$ bits.

(b) $H(X, Y) = -\sum_{x,y} p(x, y) \log_2 p(x, y) = 2\log 3 - 4/9$.

(c) In order to find $H(X|Y)$, we need to find $p(x|y)$, which is given by $p(x|y) = p(x, y)/p(y)$.
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.

Alternatively, using the results of (a) and (b), we directly compute $H(Y|X) = H(X, Y) - H(X) = \log 3 - 4/9 = H(Y|X)$.

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

(e) Cf. lecture notes or the Wikipedia page on mutual information[1].

Exercise 2.

(a) By using the chain rule, $H(X_1,X_2,...,X_k) = \sum_{i=1}^k H(X_i|X_{i-1},....,X_1)$.
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, $H(X_1,X_2,...,X_k) = kH(X)$.

(b) The i-th draw is described by the random variable X_i. Since the i-th draw is independent of all previous ones, and the color of the balls drawn during the first $i-1$ draws is not known (e.g., it is forgotten; the experiment can be also described as taking $i-1$ balls from one urn and putting them into another urn without looking at them), **no information is gained** prior to the i draw. Therefore, the entropy does not change with i, yielding $H(X_i) = H(X)$, where X stands for the color of the ball at an arbitrary draw.

(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.

To prove this, let the total number of balls in the urn be $t = r + g + b$. Then model the experiment by a tree where each level represents a draw and each branch is labeled by a particular color. For example, the probability that the first ball drawn is red is $p_r = \frac{r}{t}$, and the second ball drawn is green is $p_g = \frac{g}{t-1}$. Now if the order of the balls drawn is reversed, the probabilities become $p_g = \frac{g}{t}$ and $p_r = \frac{r}{t-1}$, respectively. However, the product of the two probabilities remain the same:

$$\frac{r}{t} \cdot \frac{g}{t-1} = \frac{r}{t-1} \cdot \frac{g}{t}$$

This reasoning can be used for any path in the tree, proving the relation.

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

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

since getting a red ball for the second draw may be preceded by drawing a red, green or blue ball first. By using the result of (c), we have

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

[1] http://en.wikipedia.org/wiki/Mutual_information

(e) The previous result shows that $p(X_2 = r) = p(X_1 = r)$. Similarly, $p(X_2 = g) = p(X_1 = g)$ and $p(X_2 = b) = p(X_1 = b)$.

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

The results of (e) and (f) can be trivially generalized for the subsequent draws: $p(X_1 = c_i) = p(X_2 = c_i) = \cdots = p(X_k = c_i)$, yielding $H(X_1) = H(X_2) = \cdots = H(X_k)$, what constitutes the constructive proof of (b).

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

Exercise 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 information is given by

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

Thus, $I(X;Y,Z) = I(Y,Z;X) = I(Y;X) + I(Z;X|Y)$ and $I(Y,Z;X) = I(Z;X) + I(Y;X|Z)$.
Furthermore, we have the definition (see lecture)

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

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

(c) Using the result of (b), $I(X;Z) \leq I(X;Y) = H(Y) - H(Y|X)$. Now $\max\{I(X;Y)\} = \log k$ 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 $I(X;Z) \leq \log k$.

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

Exercise 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: $p(x_1,x_2,...x_n) = p^k(1-p)^{n-k} \approx p^{np}(1-p)^{n(1-p)}$.
The latter can be 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.

Exercise 5.

(a) By replacing $H(Y|X) = H(X,Y) - H(X)$ in the definition of the distance, we obtain $\rho(X,Y) = 2H(X,Y) - H(X) - H(Y)$. Furthermore, the definition $I(X;Y) = H(X) + H(Y) - H(X,Y)$ gives us the second expression.

(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) we get $A = 2[H(X,Y) + H(Y,Z) - H(Y) - H(X,Z)]$. Using the strong subadditivity $H(X,Y) + H(Y,Z) - H(Y) \geq H(X,Y,Z))$, we have $A \geq 2[H(X,Y,Z) - H(X,Z)] \equiv 2H(Y|X,Z) \geq 0$.

Exercise 6.

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

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

		$Y =$		
P(X,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 $I(X;Y) = 0$ bit since X and Y are independent and thus $H(Y|X) = H(Y)$.
Furthermore, $I(X;Y|Z) = H(X|Z) - H(X|Y,Z)$. In our example X is fixed if one knows Y and Z. Thus, $H(X|Y,Z) = 0$. This implies $I(X;Y|Z) = H(X|Z)$. One obtains $I(X;Y|Z) = 1$ bit. One verifies that $I(X;Y;Z) = -1$ bit < 0 bit and $I(X;Y|Z) > I(X;Y)$. We confirm furthermore, that $I(X;Z) = I(Y;Z) = 0$. 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 $I(X;Y;Z)$.

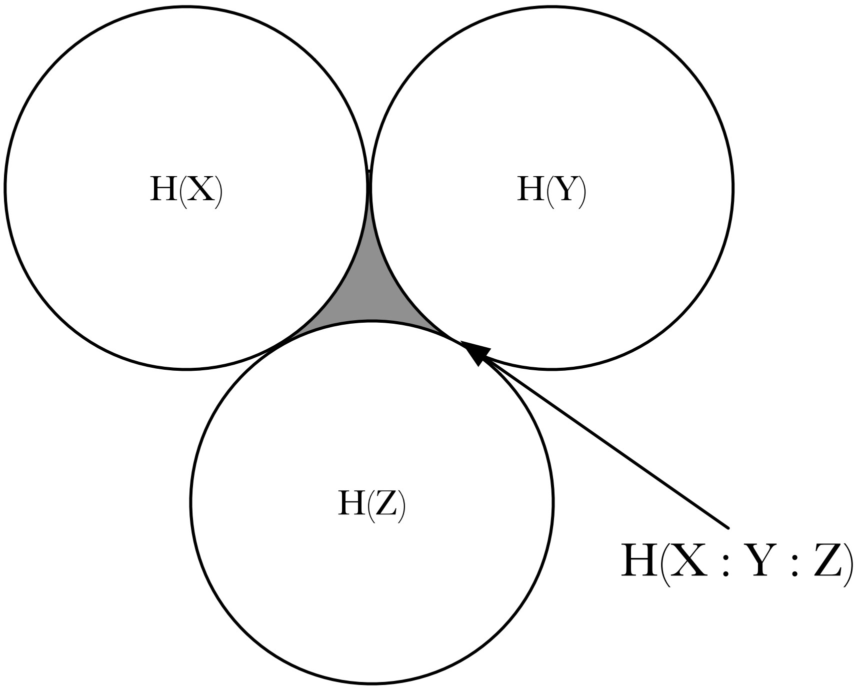

Figure 1: Venn diagram depicting the example of the Exercise 6(b).