Information Theory

Exercise Sheet 1

Recall that a discrete distribution is defined as $(p_i)_{i\in\mathcal{I}}$ where $\mathcal{I} \subseteq \mathbb{N}$, $p_i \geq 0$ for any $i \in \mathcal{I}$ and $\sum_{i\in\mathcal{I}} p_i = 1$.

Exercise 1. *Given two discrete random variable* $\mathbf{X}, \mathbf{Y} : \Omega \to \mathcal{X}$*,*

- *is* $\Big(\mathbb{P}\left(\mathbf{X} = x \mid \mathbf{Y} = y\right)\Big)_{x\in\mathcal{X}}$ *a distribution for* $y \in \mathcal{X}$ *such that* $\mathbb{P}(\mathbf{Y} = y) > 0$*?*
- *is* $\Big(\mathbb{P}\left(\mathbf{X} = x \mid \mathbf{Y} = y\right)\Big)_{y\in\mathcal{X}}$ *a distribution for* $x \in \mathcal{X}$*?*

The quantity $\mathbb{P}\left(\mathbf{X} = x \mid \mathbf{Y} = y\right)$ *is sometimes called the likelihood of* y*.*

Exercise 2. *There are eleven urns labelled by* $u \in \{0, \ldots, 10\}$ *each containing ten balls. Urn* u *contains* u *black balls and* $10 - u$ *white balls.*

1. *Alice selects an urn* u *at random and draws* N *times with replacement from that urn, obtaining* n_B *blacks and* $N - n_B$ *whites. Alice's friend, Bob, looks on. If after* N *draws* n_B *blacks have been drawn, what is the probability as function of* N*,* u *and* n_B *that the urn Alice is using is urn* u*, from Bob's point of view? (Bob doesn't know the value of* u*.)*
2. *Assuming again that Bob has observed* n_B *blacks in* N *draws, let Alice draw another ball from the same urn. What is the probability that the next drawn ball is a black?*

Exercise 3 (The likelihood principle)**.** *Urn A contains three balls: one black, and two white; urn B contains three balls: two black, and one white. One of the urns is selected at random and one ball is drawn. The ball is black. What is the probability that the selected urn is urn A?*

Exercise 4 (Entropy and conditional entropy)**.** *Given two random variables* $\mathbf{X}$ *and* $\mathbf{Y}$*, prove that*

$$H(\mathbf{X} \mid \mathbf{Y}) + H(\mathbf{Y}) = H(\mathbf{X}, \mathbf{Y})$$

Exercise 5 (Entropy: chain rule)**.** *Let* $\mathbf{Y}, \mathbf{X}_1 \ldots, \mathbf{X}_L : \Omega \to \mathcal{X}$ *be random variables. Using definitions of the lecture:*

$$H(\mathbf{X}_1, \ldots, \mathbf{X}_L) = -\sum_{x_1,\ldots,x_L} p(x_1, \ldots, x_L) \log_2 p(x_1, \ldots, x_L)$$

$$H\left(\mathbf{Y} \mid \mathbf{X}_1, \ldots, \mathbf{X}_L\right) = -\sum_{y,x_1,\ldots,x_L} p(y, x_1, \ldots, x_L) \log_2 p(y \mid x_1, \ldots, x_L)$$

1. **Independent Case:** *show that when the* $\mathbf{X}_i$*'s are independent,*

$$H(\mathbf{X}_1, \ldots, \mathbf{X}_L) = \sum_{i=1}^{L} H(\mathbf{X}_i)$$

2. **General case:** *show that the following (known as the entropy chain rule formula),*

$$H(\mathbf{X}_1, \ldots, \mathbf{X}_L) = \sum_{i=1}^{L} H(\mathbf{X}_i \mid \mathbf{X}_{i-1}, \ldots, \mathbf{X}_1)$$

Exercise 6. *Let* $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \in \{0,1\}^3$ *be random variables such that*

$$p_{\mathbf{XYZ}}(0,0,0) = \frac{1}{4} \quad \textit{and} \quad p_{\mathbf{XYZ}}(0,1,0) = \frac{1}{4}$$

$$p_{\mathbf{XYZ}}(1,0,0) = \frac{1}{4} \quad \textit{and} \quad p_{\mathbf{XYZ}}(1,0,1) = \frac{1}{4}$$

Compute $H(\mathbf{X})$*,* $H(\mathbf{Y} \mid \mathbf{X})$*,* $H(\mathbf{Z}|\mathbf{X}, \mathbf{Y})$*. Find* $H(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ *with the following computations: direct application of the definition and chain-rule.*

Recall that $h(1/4) \approx 0.811$*. Compute* $H(\mathbf{Y})$*. Check that* $H(\mathbf{Y} \mid \mathbf{X}) \leq H(\mathbf{Y})$*. What does* $\mathbf{X}$ *bring as information on* $\mathbf{Y}$ *and reciprocally?*

Exercise 7. *Show that* $H(\mathbf{Y} \mid \mathbf{X}) = 0$ *if and only if* $\mathbf{Y}$ *is a function of* $\mathbf{X}$*, i.e., for any* x *such that* $p(x) > 0$*, then it exists* y *such that* $p(y \mid x) = 1$*.*

To solve the following exercise, we will admit the following result (known as *strong additive theorem*), for any random variable $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$,

$$H(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) - H(\mathbf{Y}, \mathbf{Z}) \leq H(\mathbf{X}, \mathbf{Y}) - H(\mathbf{Y})$$

Exercise 8 (Data processing inequality, once information has been lost, it is gone forever)**.**

In many applications of interest we perform computations on the information we have available, but that information si imperfect, it has been subjected to some noise before it becomes available to us. A basic inequality of information theory, the data processing inequality, states that information about the output of a source can only decrease with time: once information has been lost, it is gone forever. Making this statement more precise is the goal of this exercise.

The intuitive notion of information processing is captured in the idea of a Markov chain of random variables. A Markov chain (of order 1*) is a sequence* $\mathbf{X}_1 \to \mathbf{X}_2 \to \cdots$ *of random variable such that*

$$\forall n \in \mathbb{N},\ \mathbb{P}\left(\mathbf{X}_{n+1} = x_n \mid \mathbf{X}_n = x_n, \ldots, \mathbf{X}_1 = x_1\right) = \mathbb{P}(\mathbf{X}_{n+1} = x_{n+1} \mid \mathbf{X}_n = x_n)$$

Let $\mathbf{X} \to \mathbf{Y} \to \mathbf{Z}$ *be a Markov Chain.*

1. *Show that* $\mathbf{Z} \to \mathbf{Y} \to \mathbf{X}$ *is a Markov Chain.*
2. *Show that,*
$$H(\mathbf{X}) \geq I(\mathbf{X}, \mathbf{Y}) \geq I(\mathbf{X}, \mathbf{Z})$$
How do you interpret this result?
3. *Let* $g : \mathcal{X} \to \mathcal{X}$*. Show that,*
$$I(\mathbf{X}, \mathbf{Y}) \geq I(\mathbf{X}, g(\mathbf{Y}))$$
How do you interpret this result?

Exercise 9 (Gibb's inequality and first consequences!)**.**

1. *Recall that for all* $x \geq 0$*,*
$$\ln x \leq x - 1$$
with equality if and only if $x = 1$*.*
Deduce Gibb's inequality
$$D_{\mathrm{KL}}(\mathbf{X}||\mathbf{Y}) \geq 0$$
with equality if and only if $\mathbf{X} = \mathbf{Y}$*.*
2. $H(\mathbf{X}) \leq \log_2 \sharp\mathcal{X}$ *with equality if and only if* $\mathbf{X}$ *is the uniform distribution over* $\mathcal{X}$*.*

3. *$H(\mathbf{X}, \mathbf{Y}) \leq H(\mathbf{X}) + H(\mathbf{Y})$ with equality if and only $\mathbf{X}$ and $\mathbf{Y}$ are independent.*

Exercise 10. *Show that,*

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

You can admit that conditioning decreases the entropy, i.e. $H(\mathbf{X} \mid \mathbf{Y}) \leq H(\mathbf{X})$.

Exercise 11 (The password problem)**.** *Someone (probably a malicious person) tries to access to some service protected by a password which is unknown. Let $\mathcal{M} = \{0,1\}^n$ be the set of possible passwords.*

Let us suppose that the protected service is perfect, i.e., the only possibility for the attacker is to try passwords one by one. Let us also suppose that the secret password was chosen according to some random variable with entropy h. Let us denote by decreasing order the probability p_i that the word m_i was chosen according to this distribution.

1. *Show that the best strategy consists in testing words one by one in the order given by the probabilities p_i. Give the expected number $\mathcal{N}(p)$ as function of the p_i's.*

2. *Let $p = (p_i)_{i\geq 1}$ and $q = (q_i)_{i\geq 1}$ be two probability distributions such that series p_i et q_i are in decreasing order. Suppose that,*

   $$q_i = (1-\alpha)\alpha^{i-1}$$

   for some $0 < \alpha < 1$. Show that if $H(p) = H(q)$, then

   $$\mathcal{N}(p) \geq \mathcal{N}(q)$$

3. *Compute the entropy $H(q)$ as function of α. Let H_α be this quantity. Recall that,*

   $$\sum_{i\geq 1} \alpha^{i-1} = 1/(1-\alpha) \quad \text{and} \quad \sum_{i\geq 1} i\alpha^{i-1} = 1/(1-\alpha)^2$$

4. *Deduce that for all $0 < \alpha < 1$ we have $1 < (1-\alpha)2^{H_\alpha} < e$.*

5. *Deduce that $\mathcal{N}(p) > \frac{1}{e}2^h$. Interpret this result.*