Exercise 2.1 Upper bound on entropy [EE5139]

In the lecture notes we show that $H(X) \leq \log |\mathcal{X}|$ for binary random variables. Show this statement for general discrete random variables on any (finite) alphabet $\mathcal{X}$.
Solution: Let us define $f(t) = -t \log t$, which is strictly concave in t.

$$\begin{aligned} H(X) &= -\sum_{x \in \mathcal{X}} p(x) \log p(x) \\ &= \sum_{x \in \mathcal{X}} f(p(x)) \\ &= |\mathcal{X}| \sum_{x \in \mathcal{X}} \frac{1}{|\mathcal{X}|} f(p(x)) \\ &\leq |\mathcal{X}| f\left(\sum_{x \in \mathcal{X}} \frac{1}{|\mathcal{X}|} p(x) \right) \\ &= |\mathcal{X}| f\left(\frac{1}{|\mathcal{X}|} \right) \\ &= \log |\mathcal{X}|, \end{aligned}$$

where the inequality follows from the concavity of f.

Exercise 2.2 Relative entropy as a parent quantity [all]

Let X and Y be random variables on alphabets $\mathcal{X}$ and $\mathcal{Y}$ with joint pmf P_{XY}. Moreover, let U be a uniform random variable on $\mathcal{X}$. Show the following relations:

a.) $H(X) = \log |\mathcal{X}| - D(P_X \| U_X)$.

Solution:

$$H(X) = -\sum_{x \in \mathcal{X}} P_X(x) \log P_X(x) = -\sum_{x \in \mathcal{X}} P_X(x) \left(\log \frac{P_X(x)}{1/|\mathcal{X}|} + \log \frac{1}{|\mathcal{X}|} \right) = \log |\mathcal{X}| - D(P_X \| U_X).$$

b.) $H(X|Y) = \log |\mathcal{X}| - D(P_{XY} \| U_X \times P_Y)$.

Solution:

$$\begin{aligned} H(X|Y) &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_{XY}(x,y) \log P_{X|Y}(x|y) \\ &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_{XY}(x,y) \left(\log \frac{P_{XY}(x,y)}{P_Y(y)/|\mathcal{X}|} + \log \frac{1}{|\mathcal{X}|} \right) \\ &= \log |\mathcal{X}| - D(P_{XY} \| U_X \times P_Y). \end{aligned}$$

c.) $I(X : Y) = D(P_{XY} \| P_X \times P_Y)$.

Solution:

$$
\begin{aligned}
I(X:Y) = H(Y) - H(Y|X) &= \sum_{y\in\mathcal{Y}} P_Y(y)\log\frac{1}{P_Y(y)} - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{XY}(x,y)\log\frac{P_X(x)}{P_{XY}(x,y)} \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{XY}(x,y)\log\frac{1}{P_Y(y)} - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{XY}(x,y)\log\frac{P_X(x)}{P_{XY}(x,y)} \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{XY}(x,y)\log\frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} \\
&= D(P_{XY}\|P_X\times P_Y).
\end{aligned}
$$

Exercise 2.3 Example correlations [EE5139]

For each item, find an example of random variables X, Y and Z (you can restrict the alphabet size to at most 2 bits) such that the desired relations holds:

a.) $H(X|YZ) = 0$ but $H(X|Y) = H(X|Z) = 1$.

Solution: We may choose binary random variables satisfying $X = Y \oplus Z$ with Y and Z uniform and independent.

b.) $I(X:Y|Z) = 1$ but $I(X:Y) = 0$.

Solution: We may choose X and Y uniform and independent with $Z = X \oplus Y$.

c.) $I(X:Y) = 1$ but $I(X:Y|Z) = 0$.

Solution: We may choose binary $X = Y = Z$ uniform.

d.) $I(X:Y) = I(X:Z) = 1$ but $I(Y:Z) = 0$.

Solution: We may choose $X = (Y,Z)$ with Y and Z independent and uniform.

Exercise 2.4 Information spectrum [EE6139]

Given a random variable X governed by the pmf P or an alternative pmf Q, the log-likelihood ratio is defined as the random variable $Z(X) = \log\frac{P(X)}{Q(X)}$.

a.) We have seen that the expectation value of Z (under P) is the relative entropy

$$
\mathbb{E}[Z] = \sum_{x\in\mathcal{X}} P(x)\log\frac{P(x)}{Q(x)} = D(P\|Q)\,. \tag{1}
$$

Give an expression for $\mathrm{Var}[Z]$ (under P). This quantity is called the relative entropy variance and denoted by $V(P\|Q)$.

Solution:

$$
V(P\|Q) = \mathrm{Var}[Z] = \sum_{x\in\mathcal{X}} P(x)\bigg(\log\frac{P(x)}{Q(x)} - D(P\|Q)\bigg)^2.
$$

Consider now a sequence of i.i.d. random variables $X^n = (X_1, X_2, \ldots, X_n)$ on $\mathcal{X}^n$ where each X_i is governed by the pmf P or an alternative pmf Q. We are interested in pmf of the log-likelihood ratio $Z(X^n)$.

b.) Show that $Z(X^n) = \sum_{i=1}^n Z(X_i)$. What is $\mathbb{E}[Z^n]$ and $\mathrm{Var}[Z^n]$?

Solution:

$$Z(X^n) = \log \frac{P(X^n)}{Q(X^n)} = \log \frac{\prod_{i=1}^n P(X_i)}{\prod_{i=1}^n Q(X_i)} = \sum_{i=1}^n \log \frac{P(X)}{Q(X)} = \sum_{i=1}^n Z(X_i).$$

$$\mathbb{E}[Z^n] = \sum_{i=1}^n \mathbb{E}[Z(X_i)] = nD(P\|Q).$$

We use independence to write

$$\mathrm{Var}[Z^n] = \sum_{i=1}^n \mathrm{Var}[Z(X_i)] = nV(P\|Q).$$

c.) Let us now consider the quantity $\Pr[Z(X^n) \le nR]$ in the limit of large n for different values of R. Show that

$$\lim_{n\to\infty} \Pr[Z(X^n) \le nR] = \begin{cases} 0 & \text{if} \quad R < D(P\|Q) \\ 1 & \text{if} \quad R > D(P\|Q) \end{cases}. \tag{2}$$

Hint: Argue using the weak law of large numbers.

Solution: Using the weak law of large numbers, we have for any positive number $\epsilon > 0$,

$$\lim_{n\to\infty} \Pr\left[\left|\frac{\sum_{i=1}^n Z(X_i)}{n} - D(P\|Q)\right| > \epsilon\right] = 0,$$

and

$$\lim_{n\to\infty} \Pr\left[\left|\frac{\sum_{i=1}^n Z(X_i)}{n} - D(P\|Q)\right| \le \epsilon\right] = 1.$$

Then, in particular,

$$\lim_{n\to\infty} \Pr[Z(X^n) \le n(D(P\|Q) - \epsilon)] = 0,$$
$$\lim_{n\to\infty} \Pr[Z(X^n) \le n(D(P\|Q) + \epsilon)] = 1.$$

Now, if $R < D(P\|Q)$ then there also exists an $\epsilon > 0$ such that $R \le D(P\|Q) - \epsilon$. And similalry, if $R > D(P\|Q)$ then there exists an $\epsilon > 0$ such that $R \ge D(P\|Q) + \epsilon$. Hence, the above inequalities imply the desired result.

d.) Later on in the lecture we will encounter the quantity

$$D_s^\epsilon(P^n\|Q^n) := \sup\{k \in \mathbb{R} : \Pr[Z(X^n) \le k] \le \epsilon\}, \tag{3}$$

which, in words, is asking the largest k such that the tail of the distribution of Z that lies below k has cumulative probability at most ϵ. Show that $D_s^\epsilon(P^n\|Q^n) = nD(P\|Q) + o(n)$, or equivalently,

$$\lim_{n\to\infty} \frac{1}{n} D_s^\epsilon(P^n\|Q^n) = D(P\|Q)\,. \tag{4}$$

Hint: Verify that $\frac{1}{n} D_s^\epsilon(P^n\|Q^n) = \sup\{k \in \mathbb{R} : \Pr[\frac{1}{n} Z(X^n) \le k] \le \epsilon\}$.

Solution: From the previous item and the definition of the limit we know that for any $R < D(P\|Q)$ and $\epsilon > 0$, there exists an $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$ we have

$$\Pr[Z(X^n) \leq nR] \leq \epsilon\,. \tag{5}$$

This implies that in the definition of (3) we are allowed to choose any $k \leq nR$ and thus by taking the supremum we get

$$D_s^\epsilon(P^n\|Q^n) \geq \sup\{k \in \mathbb{R} : k < nR\} = nR\,. \tag{6}$$

Taking the limit $n \to \infty$ yields the desired lower bound, for all $R < D(P\|Q)$,

$$\lim_{n\to\infty} \frac{1}{n} D_s^\epsilon(P^n\|Q^n) \geq R, \tag{7}$$

And hence $\lim_{n\to\infty} \frac{1}{n} D_s^\epsilon(P^n\|Q^n) \geq D(P\|Q)$ since this holds for all $R < D(P\|Q)$.

We can argue similarly in the opposite direction. For any $R > D(P\|Q)$ and $\mu > 0$, there exists an $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$ we have

$$\Pr[Z(X^n) \leq nR] \geq 1 - \mu\,. \tag{8}$$

If we choose μ small enough so that $1 - \mu > \epsilon$ then this implies that any $k \geq nR$ violates the constraint on the probability in the definition of $D_s^\epsilon(P^n\|Q^n)$, and thus we must have

$$D_s^\epsilon(P^n\|Q^n) \leq nR \tag{9}$$

Taking again the limit we find that

$$\lim_{n\to\infty} \frac{1}{n} D_s^\epsilon(P^n\|Q^n) \leq R, \tag{10}$$

Since this holds for any $R > D(P\|Q)$ we deduce that $\lim_{n\to\infty} \frac{1}{n} D_s^\epsilon(P^n\|Q^n) \leq D(P\|Q)$.

e.) Optional: Show that in the next order in n, we have

$$D_s^\epsilon(P^n\|Q^n) = nD(P\|Q) + \sqrt{nV(P\|Q)}\,\Phi^{-1}(\epsilon) + o\left(\sqrt{n}\right) \tag{11}$$

Can we even say something more about the $o\left(\sqrt{n}\right)$ term?

Hint: The statement can be shown using the central limit theorem. A quantitative version of the central limit theorem is the Berry-Esseen theorem. Look it up to make even stronger statements about the remainder term.

Solution: We give here actually an even stronger bound, using the Berry–Eseen theorem, which tells us how quickly the renormalised distribution approaches the Gaussian distribution in the central limit theorem. To make this precise we define

$$T(P\|Q) := \sum_{x\in\mathcal{X}} P(x) \left| \log \frac{P(x)}{Q(x)} - D(P\|Q) \right|^3.$$

By the Berry-Esseen theorem, we have

$$\left| \Pr[Z(X^n) \leq k] - \Phi\left(\frac{\sqrt{n}(k/n - D(P\|Q))}{\sqrt{V(P\|Q)}} \right) \right| \leq \frac{6T(P\|Q)}{\sqrt{nV^3(P\|Q)}}.$$

By letting

$$\Phi\left(\frac{\sqrt{n}(k/n - D(P\|Q))}{\sqrt{V(P\|Q)}}\right) + \frac{6T(P\|Q)}{\sqrt{nV^3(P\|Q)}} \geq \epsilon$$

and

$$\Phi\left(\frac{\sqrt{n}(k/n - D(P\|Q))}{\sqrt{V(P\|Q)}}\right) - \frac{6T(P\|Q)}{\sqrt{nV^3(P\|Q)}} \leq \epsilon,$$

we can constrain k as follows:

$$\begin{aligned} &nD(P\|Q) + \sqrt{nV(P\|Q)}\Phi^{-1}\left(\epsilon - \frac{6T(P\|Q)}{\sqrt{nV^3(P\|Q)}}\right) \\ &\leq k \\ &\leq nD(P\|Q) + \sqrt{nV(P\|Q)}\Phi^{-1}\left(\epsilon + \frac{6T(P\|Q)}{\sqrt{nV^3(P\|Q)}}\right). \end{aligned}$$

This implies the two bounds

$$D_s^\epsilon(P^n\|Q^n) \geq nD(P\|Q) + \sqrt{nV(P\|Q)}\Phi^{-1}\left(\epsilon - \frac{6T(P\|Q)}{\sqrt{nV^3(P\|Q)}}\right) \tag{12}$$

$$D_s^\epsilon(P^n\|Q^n) \leq nD(P\|Q) + \sqrt{nV(P\|Q)}\Phi^{-1}\left(\epsilon + \frac{6T(P\|Q)}{\sqrt{nV^3(P\|Q)}}\right) \tag{13}$$

Equivalently,

$$\lim_{n\to\infty} \frac{1}{n} D_s^\epsilon(P^n\|Q^n) = D(P\|Q)\,. \tag{14}$$

If $V(P\|Q) > 0$ and $T(P\|Q) < \infty$, the term $\frac{6T(P\|Q)}{\sqrt{nV^3(P\|Q)}}$ is equal to $c/\sqrt{n}$ for some finite $c > 0$. By Taylor expansions,

$$\Phi^{-1}\left(\epsilon \pm \frac{c}{\sqrt{n}}\right) = \Phi^{-1}(\epsilon) + O\left(\frac{1}{\sqrt{n}}\right).$$

By plugging in the Taylor expansion, we can get the result.

Exercise 2.5 Independence and mutual information [all]

Consider two sequences of random variables $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$. Show that if $X_1, \ldots, X_n$ are mutually independent, then

$$I(X_1, \ldots, X_n : Y_1, \ldots, Y_n) \geq \sum_{i=1}^n I(X_i : Y_i)$$

while if given Y_i the random variable X_i is conditionally independent of all the remaining random variables for all $i = 1, \ldots, n$, then

$$I(X_1, \ldots, X_n : Y_1, \ldots, Y_n) \leq \sum_{i=1}^n I(X_i : Y_i)$$

Solution: For the first claim, consider

$$\begin{aligned}
I(X_1,\ldots,X_n : Y_1,\ldots,Y_n) &= \sum_{i=1}^{n} I(X_i : Y_1,\ldots,Y_n | X_1,\ldots,X_{i-1}) \\
&= \sum_{i=1}^{n} I(X_i : Y_1,\ldots,Y_n, X_1,\ldots,X_{i-1}) - I(X_i : X_1,\ldots,X_{i-1}) \\
&= \sum_{i=1}^{n} I(X_i : Y_1,\ldots,Y_n, X_1,\ldots,X_{i-1}) \\
&\geq \sum_{i=1}^{n} I(X_i : Y_i)
\end{aligned}$$

where the third equality is by independence.
For the second claim, consider

$$\begin{aligned}
I(X_1,\ldots,X_n : Y_1,\ldots,Y_n) &= H(X_1,\ldots,X_n) - H(X_1,\ldots,X_n | Y_1,\ldots,Y_n) \\
&= H(X_1,\ldots,X_n) - \sum_{i=1}^{n} H(X_i | Y_1,\ldots,Y_n, X_1,\ldots,X_{i-1}) \\
&= H(X_1,\ldots,X_n) - \sum_{i=1}^{n} H(X_i | Y_i) \\
&\leq \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i | Y_i) \\
&= \sum_{i=1}^{n} I(X_i : Y_i)
\end{aligned}$$

where the third equality is by the fact that given Y_i, X_i is conditionally independent of all other random variables for $i = 1,\ldots,n$ so

$$H(X_i | Y_1,\ldots,Y_n, X_1,\ldots,X_{i-1}) = H(X_i | Y_i).$$