Information Theory
University of Amsterdam, 2019
teachers: Christian Schaffner, Yfke Dulek TAs: Esteban Landerreche, Maximilian Siemers

Team Name: Hadamard
Student Names: Leonardo Romor, Maurice Frank, Simone Astarita, XinYu Fu, Yije Zhang

Homework set 2

Contents

Unless otherwise stated, you should provide exact answers rather than rounded numbers (e.g., $\log 3$ instead of 1.585) for non-programming exercises.

Problem 1 (5 pt): Entropy of functions of a random variable

Let X be a random variable, and let f be a function of X.

(2 pt) **a.** Show that $H(f(X)|X) = 0$.

Solution: ... □

(3 pt) **b.** Show that $H(f(X)) \leq H(X)$.
Hint: use the chain rule.

Solution: ... □

Problem 2 (8 pt): Relative and cross entropy

(5 pt) **a.** Prove that for any two distributions P and Q over $\mathcal{X}$, $D(P||Q) \geq 0$, and that equality holds if and only if $P = Q$.

Solution: We begin by showing that $D(P||Q) \geq 0$.
Proof. By definition of relative entropy and by exploiting Jensen inequality for the convex function $-\log(x)$ we have that:

$$\begin{aligned} D(P||Q) &= \mathbb{E}_{x\sim P}\left[-\log(\frac{Q(x)}{P(x)})\right] \\ &\geq -\log\left(\mathbb{E}_{x\sim P}\left[\frac{Q(x)}{P(x)}\right]\right) \\ &= -log\left[\sum_x P(x)\frac{Q(x)}{P(x)}\right] = \log 1 = 0 \end{aligned} \tag{1}$$

hence the thesis. □

For the second proof we use again the Jensen inequality and the statement that the Jensen inequality becomes an equality if and only if all the elements inside the convex function are equal.

Proof. By the axioms of probability, we have that $\sum_{x\in\mathcal{X}} P(x) = 1$. Without loss of generality, we can assume that we labeled each element of $\mathcal{X}$ with an positive integer $i \in \mathbb{N}$.

The Jensen inequality becomes an equality if and only if the following condition holds:

$$\frac{P(x_1)}{Q(x_1)} = \cdots = \frac{P(x_n)}{Q(x_n)} \tag{2}$$

Since *relative entropy* $D(P||Q)$ is defined only if $Q(x) \neq 0$ if $P(x) \neq 0$. We can then write for any $P(x_i)$ and $P(x_j)$:

$$P(x_j) = \frac{P(x_i)Q(x_j)}{Q(x_i)} \tag{3}$$

but then we can write, $\forall x_i$:

$$\begin{aligned} \sum_j P(x_j) &= 1 \\ P(x_i) + \sum_{j\neq i} \frac{P(x_i)Q(x_j)}{Q(x_i)} &= 1 \\ P(x_i) + \frac{P(x_i)}{Q(x_i)}(1 - Q(x_i)) &= 1 \\ P(x_i) &= Q(x_i) \end{aligned} \tag{4}$$

Hence $D(P||Q) = 0 \iff P(x) = Q(x), \forall x \in \mathcal{X}$. □

□

(1 pt) **b.** We have seen that the mutual information can be expressed in terms of the relative entropy, i.e. that $I(X;Y) = D(P_{XY}||P_X \cdot P_Y)$. Use (a) and this fact to show that $H(X|Y) \leq H(X)$.

Solution:

Proof. By definition and since $D(P||Q) \geq 0$:

$$H(X) = H(X|Y) + I(X;Y) = H(X|Y) + D(P_{XY}||P_X P_Y) \geq H(X|Y) \tag{5}$$

□

□

(2 pt) **c.** We have seen a relation between relative entropy and cross entropy in Intro/Team Quiz 02. Use this relation to express the mutual information $I(X;Y)$ in terms of Shannon entropies of X and Y (such as $H(X), H(Y), H(X|Y), H(Y|X), H(YX)$) and of the cross entropy H_c of P_{XY} and $P_X \cdot P_Y$.

Solution:

By the previous exercise we know that $I(X;Y) = D(P_{XY}||P_x P_Y)$. From the previous quiz, we also know that $H_c(P,Q) = H(P) + D(P||Q)$. If we use the previous equalities rearrange the equations we get:

$$I(X;Y) = D(P_{XY}||P_x P_Y) = H_c(P_{XY}, P_x P_Y) - H(P_{XY}) \tag{6}$$

hence we have expressed $I(X;Y)$ in terms of the cross entropy P_{XY}, $P_x P_Y$ and the joint entropy $H(X,Y)$.

□

Problem 3 (8 pt): Programming

Note: You do not have to hand in your code. As before, describe briefly what you did to reach your answer, what choices you made, and how you treated edge cases if those arose.

(1 pt) **a.** In the first programming quiz for this module, you computed the entropy of sampling a single letter from the story of Jack and the beanstalk.

If you instead decided to sample a single *word* from the same story, do you expect the entropy to be higher or lower compared to sampling a single letter? Why?

Solution: ... □

(3 pt) **b.** Verify your answer from (a) by writing a program that computes the entropy explicitly. Use the same text file as input.

Solution: ... □

(1 pt) **c.** In the second programming quiz for this module, you computed the entropy of the second letter given the first when sampling a pair of letters from the story of Jack and the beanstalk. Let P_{XY} denote the joint distribution where X is the first and Y is the second letter (again negleticing the beginning and end of the text).

Explain in words why $P_X = P_Y$.

Solution: ... □

(2 pt) **d.** For P_{XY} the bigram distribution from above, write a program to compute the cross entropy $H_C(P_{XY}, P_X \cdot P_Y)$ between the bigram distribution and the independent single-letter distribution.

Solution: ... □

(1 pt) **e.** Express in words what this quantity $H_C(P_{XY}, P_X \cdot P_Y)$ means.

Solution: ... □