
Problem Set 2 Information Measures

For the Exercise Sessions on Sept 15 and 22— **Due: Tue, September 29, 10am, on Moodle**

1 The Homework

The following four problems are the core of this homework. You work out solutions and turn them in. Problem 1 is suitable for Sept 15. Problem 2 can be tackled on Sept 15, and on Sept 22, you'll have even more tools to address it. Problems 3 and 4 are suitable for Sept 22.

Problem 1: Entropy and Geometry

Suppose X , Y and Z are random variables.

- (a) Show that $H(X) + H(Y) + H(Z) \geq \frac{1}{2}[H(X, Y) + H(Y, Z) + H(Z, X)]$.
- (b) Show that $H(X, Y) + H(Y, Z) \geq H(X, Y, Z) + H(Y)$.

Solution 1. (a) By the sub-additivity of Entropy we know that

$$\begin{aligned} H(X, Y) &\leq H(X) + H(Y) \\ H(Y, Z) &\leq H(Y) + H(Z) \\ H(X, Z) &\leq H(X) + H(Z). \end{aligned}$$

Adding the three inequalities together we retrieve:

$$H(X) + H(Y) + H(Z) \geq \frac{1}{2} (H(X, Y) + H(Y, Z) + H(Z, X)).$$

(b) It is easier to show

$$H(X, Y) + H(Y, Z) - (H(X, Y, Z) + H(Y)) \geq 0.$$

Indeed we have that:

$$H(X|Y) - H(X|Y, Z) = I(X; Z|Y) \geq 0.$$

Problem 2: DPI for TV distance

Show that the Total Variation Distance satisfies the Data Processing Inequality. That is, using the definitions given in Lemma 4.3 of the Lecture Notes: a probability kernel $W(y|x)$, an original distribution P , and the distribution $\tilde{P}$ whose probability mass function is $\tilde{p}(y) = \sum_x W(y|x)p(x)$, show that

$$\delta(\tilde{P}, \tilde{Q}) \leq \delta(P, Q),$$

where $\delta(P, Q)$ denotes the Total Variation distance between P and Q .

Solution 2. There are several ways in which you can proceed. We provide three different solutions — challenge: can you come up with even more?

Proof 1. Via the variational representation of TV distance, which is Lemma 4.10 of the lecture notes.

$$\delta(\tilde{P}, \tilde{Q}) = \max_{f: \mathcal{Y} \rightarrow [0,1]} \mathbb{E}_{\tilde{P}}[f(Y)] - \mathbb{E}_{\tilde{Q}}[f(Y)] \quad (1)$$

$$= \max_{f: \mathcal{Y} \rightarrow [0,1]} \sum_y \tilde{p}(y)f(y) - \sum_y \tilde{q}(y)f(y) \quad (2)$$

$$= \max_{f: \mathcal{Y} \rightarrow [0,1]} \sum_y \left(\sum_x W(y|x)p(x) \right) f(y) - \sum_y \left(\sum_x W(y|x)q(x) \right) f(y) \quad (3)$$

$$= \max_{f: \mathcal{Y} \rightarrow [0,1]} \sum_x \sum_y W(y|x)p(x)f(y) - W(y|x)q(x)f(y) \quad (4)$$

$$= \max_{f: \mathcal{Y} \rightarrow [0,1]} \sum_x \underbrace{\left(\sum_y W(y|x)f(y) \right)}_{=\tilde{f}(x)} p(x) - \left(\sum_y W(y|x)f(y) \right) q(x) \quad (5)$$

$$= \max_{f: \mathcal{Y} \rightarrow [0,1]} \mathbb{E}_P[\tilde{f}(X)] - \mathbb{E}_Q[\tilde{f}(X)]. \quad (6)$$

Finally, we observe that if $0 \leq f(y) \leq 1$ for all y , then $0 \leq \tilde{f}(x) \leq 1$ for all x , because

$$\tilde{f}(x) = \sum_y W(y|x)f(y) \leq \sum_y W(y|x) = 1. \quad (7)$$

This is the crucial observation here. How do we exploit this observation? We will now expand the maximization to include *all* possible functions $\tilde{f}(x)$ satisfying $0 \leq \tilde{f}(x) \leq 1$ (not only the ones that are of the form $\tilde{f}(x) = \sum_y W(y|x)f(y)$ for some $f(y)$). Clearly, expanding the maximization can only *increase* the value of the maximum, that is,

$$\delta(\tilde{P}, \tilde{Q}) = \max_{f: \mathcal{Y} \rightarrow [0,1]} \mathbb{E}_P[\tilde{f}(X)] - \mathbb{E}_Q[\tilde{f}(X)] \quad (8)$$

$$\leq \max_{\tilde{f}: \mathcal{X} \rightarrow [0,1]} \mathbb{E}_P[\tilde{f}(X)] - \mathbb{E}_Q[\tilde{f}(X)] \quad (9)$$

$$= \delta(P, Q), \quad (10)$$

where in the last step, we have again used the variational representation of TV distance, which is Lemma 4.10 of the lecture notes.

Proof 2. Via the triangle inequality (that is, a “pedestrian” proof).

$$\delta(\tilde{P}, \tilde{Q}) = \frac{1}{2} \sum_y |\tilde{p}(y) - \tilde{q}(y)| \quad (11)$$

$$= \frac{1}{2} \sum_y \left| \sum_x W(y|x)p(x) - \sum_x W(y|x)q(x) \right| \quad (12)$$

$$= \frac{1}{2} \sum_y \left| \sum_x W(y|x)(p(x) - q(x)) \right| \quad (13)$$

$$\leq \frac{1}{2} \sum_y \sum_x |W(y|x)(p(x) - q(x))| \quad (14)$$

$$= \frac{1}{2} \sum_y \sum_x |W(y|x)| |p(x) - q(x)| \quad (15)$$

$$= \frac{1}{2} \sum_y \sum_x W(y|x) |p(x) - q(x)| \quad (16)$$

$$= \frac{1}{2} \sum_x \underbrace{\left(\sum_y W(y|x) \right)}_{=1} |p(x) - q(x)| \quad (17)$$

$$= \frac{1}{2} \sum_x |p(x) - q(x)| \quad (18)$$

$$= \delta(P, Q), \quad (19)$$

where the inequality step follows from the triangle inequality $|a + b| \leq |a| + |b|$.

Proof 3. Via the original definition of TV distance, Definition 4.2 in the lecture notes.

$$\delta(\tilde{P}, \tilde{Q}) = \max_{S \subseteq \mathcal{Y}} \tilde{P}(S) - \tilde{Q}(S) \quad (20)$$

$$= \max_{S \subseteq \mathcal{Y}} \sum_{y \in S} \tilde{p}(y) - \sum_{y \in S} \tilde{q}(y) \quad (21)$$

$$= \max_{S \subseteq \mathcal{Y}} \sum_{y \in S} (\tilde{p}(y) - \tilde{q}(y)) \quad (22)$$

$$= \max_{S \subseteq \mathcal{Y}} \sum_{y \in S} \left(\sum_{x \in \mathcal{X}} W(y|x)(p(x) - q(x)) \right) \quad (23)$$

$$= \max_{S \subseteq \mathcal{Y}} \sum_{x \in \mathcal{X}} \sum_{y \in S} W(y|x)(p(x) - q(x)) \quad (24)$$

Now comes the tricky step: Like in the lecture notes, let $\mathcal{A} = \{x \in \mathcal{X} : p(x) \geq q(x)\}$. Consider the last sum over $x \in \mathcal{X}$ and observe that $W(y|x) \geq 0$. Now, we note that all $x \in \mathcal{A}$ contribute non-negatively. But all $x \notin \mathcal{A}$ contribute strictly negatively to the sum. So, dropping all $x \notin \mathcal{A}$ can only make the sum

larger. That is,

$$\delta(\tilde{P}, \tilde{Q}) = \max_{S \subseteq \mathcal{Y}} \sum_{x \in \mathcal{X}} \sum_{y \in S} W(y|x)(p(x) - q(x)) \quad (25)$$

$$\leq \max_{S \subseteq \mathcal{Y}} \sum_{x \in \mathcal{A}} \sum_{y \in S} W(y|x)(p(x) - q(x)) \quad (26)$$

$$= \max_{S \subseteq \mathcal{Y}} \sum_{x \in \mathcal{A}} \left(\underbrace{\sum_{y \in S} W(y|x)}_{\leq 1} \right) (p(x) - q(x)) \quad (27)$$

$$= \max_{S \subseteq \mathcal{Y}} \sum_{x \in \mathcal{A}} (p(x) - q(x)) \quad (28)$$

$$= \sum_{x \in \mathcal{A}} (p(x) - q(x)). \quad (29)$$

To complete the proof, we recall from class that this last sum, for the very particular choice of the set $\mathcal{A}$ that we have made, is precisely the TV distance between P and Q .

Problem 3: Variational characterization of mutual information

Let X and Y be two random variables over finite alphabets $\mathcal{X}$ and $\mathcal{Y}$ with joint probability distribution P_{XY} , and let $I(X; Y)$ be their mutual information.

- (a) Show that for every function $f(X, Y)$ such that $E_{P_X P_Y}[e^{f(X, Y)}]$ is finite,

$$I(X; Y) \geq \mathbb{E}_{P_{XY}}[f(X, Y)] - \mathbb{E}_{P_Y}[\log \mathbb{E}_{P_X}[e^{f(X, Y)}]].$$

- (b) Show that there is a function $\tilde{f}(X, Y)$ such that $E_{P_X P_Y}[e^{\tilde{f}(X, Y)}]$ is finite and

$$I(X; Y) = \mathbb{E}_{P_{XY}}[\tilde{f}(X, Y)] - \mathbb{E}_{P_Y}[\log \mathbb{E}_{P_X}[e^{\tilde{f}(X, Y)}]].$$

- (c) Conclude that

$$I(X; Y) = \sup_f \mathbb{E}_{P_{XY}}[f(X, Y)] - \mathbb{E}_{P_Y}[\log \mathbb{E}_{P_X}[e^{f(X, Y)}]]$$

where the sup is over all functions f such that $E_{P_X P_Y}[e^{f(X, Y)}]$ is finite.

Solution 3. (a)

$$\begin{aligned} \mathbb{E}_{P_{XY}}[f(X, Y)] - \mathbb{E}_{P_Y}[\log \mathbb{E}_{P_X}[e^{f(X, Y)}]] &= \mathbb{E}_{P_Y}[\mathbb{E}_{P_{X|Y}}[f(X, Y)] - \log \mathbb{E}_{P_X}[e^{f(X, Y)}]] \\ &\leq \mathbb{E}_{P_Y}[D(P_{X|Y} \| P_X)] = I(X; Y) \end{aligned}$$

where the inequality is due to the Donsker-Varadhan form of the KL divergence seen in class.

(b) Pick $f(x, y) = \log \frac{P_{XY}(x, y)}{P_X(x)P_Y(y)}$. For this choice of f , $E_{P_X P_Y}[e^{f(X, Y)}]$ is finite and simple substitution shows that $\mathbb{E}_{P_{XY}}[f(X, Y)] - \mathbb{E}_{P_Y}[\log \mathbb{E}_{P_X}[e^{f(X, Y)}]] = I(X; Y)$.

(c) By (a) we know that $\sup_f \mathbb{E}_{P_{XY}}[f(X, Y)] - \mathbb{E}_{P_Y}[\log \mathbb{E}_{P_X}[e^{f(X, Y)}]]$ is a lower bound on $I(X; Y)$. By (b) we know that the bound can be achieved with $f(x, y) = \log \frac{P_{XY}(x, y)}{P_X(x)P_Y(y)}$. This proves that the bound is actually an equality.

Problem 4: Geometrical interpretation of mutual information

We saw in class that a *probability kernel* $P_{Y|X} : \mathcal{X} \rightarrow \mathcal{Y}$ is a matrix $P_{Y|X} = P_{Y|X}(y|x) : x \in \mathcal{X}, y \in \mathcal{Y}$ such that $P_{Y|X}(y|x) \geq 0$, and for each $x \in \mathcal{X}$, $\sum_y P_{Y|X}(y|x) = 1$. Let $P_X \in \Pi(\mathcal{X})$ be a probability distribution on $\mathcal{X}$. We define the *conditional KL divergence* between two probability kernels $P_{Y|X} : \mathcal{X} \rightarrow \mathcal{Y}$ and $Q_{Y|X} : \mathcal{X} \rightarrow \mathcal{Y}$ given P_X to be

$$D(P_{Y|X} \| Q_{Y|X} | P_X) \triangleq \sum_{x \in \mathcal{X}} P_X(x) D(P_{Y|X}(\cdot|x) \| Q_{Y|X}(\cdot|x))$$

where for every $x \in \mathcal{X}$, $D(P_{Y|X}(\cdot|x) \| Q_{Y|X}(\cdot|x))$ is the standard KL divergence between the two distributions $P_{Y|X}(\cdot|x)$ and $Q_{Y|X}(\cdot|x)$ over $\mathcal{Y}$. *Hint:* Recall Problem 2 above (the problem we solved together in class).

- (a) Let X and Y be two random variables with joint distribution $P_{XY} = P_X P_{Y|X}$. Show that

$$I(X; Y) = \sum_{x \in \mathcal{X}} P_X(x) D(P_{Y|X}(\cdot|x) \| P_Y)$$

where P_Y is the marginal distribution of Y . This formula shows that the mutual information can be interpreted as a weighted average of the distances between the conditional distributions $P_{Y|X}(\cdot|x)$ and the marginal distribution P_Y .

- (b) Show that for any distribution Q_Y on $\mathcal{Y}$,

$$I(X; Y) = D(P_{Y|X} \| Q_Y | P_X) - D(P_Y \| Q_Y).$$

You can think of this formula as a KL equivalent of the classical $I(X; Y) = H(Y) - H(Y|X)$.

- (c) Show that

$$I(X; Y) = \min_{Q_Y} D(P_{Y|X} \| Q_Y | P_X).$$

According to this formula, the minimizing Q_Y can be interpreted as the “center of gravity” of the conditional distributions $P_{Y|X}(\cdot|x)$, and the mutual information as its radius.

Solution 4. All the results can be proved working directly with the definitions of KL divergence and mutual information. The following is a simple solution that makes use of the results of the Problem that we solved together in class (Problem 2 on this Homework Set).

- (a)

$$I(X; Y) = D(P_X P_{Y|X} \| P_X P_Y) = D(P_{Y|X} \| P_Y | P_X) = \sum_{x \in \mathcal{X}} P_X(x) D(P_{Y|X}(\cdot|x) \| P_Y),$$

where the second equality is due to Problem 2(b).

- (b)

$$\begin{aligned} D(P_Y \| Q_Y) + I(X; Y) &= D(P_Y \| Q_Y) + D(P_{X|Y} \| P_X | P_Y) \\ &= D(P_{XY} \| P_X Q_Y) \\ &= D(P_{Y|X} \| Q_Y | P_X) \end{aligned}$$

where the first equality is due to part (a) by exchanging the roles of X and Y , the second equality is due to the chain rule of the KL divergence (Problem 2(a)), and the third equality is again due to Problem 2(b).

- (c) By part (b) we know that $I(X; Y) \leq D(P_{Y|X} \| Q_Y | P_X)$ for every Q_Y , since $D(P_Y \| Q_Y) \geq 0$. Hence, $I(X; Y) \leq \min_{Q_Y} D(P_{Y|X} \| Q_Y | P_X)$. The equality is achieved by picking $Q_Y = P_Y$, for which $D(P_{Y|X} \| Q_Y | P_X) = D(P_{Y|X} \| P_Y | P_X) = I(X; Y)$.

2 Additional Problems

For the remaining problems, you do not need to turn in solutions. For the exam, we do expect you know how to solve these problems. So, you may keep these problems around and tackle them in preparation for the Midterm and/or Final exam. Or you may solve them now:

- On Sept 15: Try Problem 5. Then try Problem 6 which includes Problem 1 above but goes beyond. Next try Problem 7. Problem 8 might interest you, but is not relevant for the exam.
- On Sept 22: Try Problem 9, it is a follow-up to Problem 4, advancing your tools for conditional KL divergence.

Problem 5: Entropy and pairwise independence

Consider three binary random variables X, Y, Z . Each of the three random variables is uniformly distributed, but they are not independent. However, we know that they are *pairwise* independent. That is, X and Y are independent, and X and Z are independent, and Y and Z are independent.

- What is $H(X, Y)$?
- Give a lower bound to the value of $H(X, Y, Z)$.
- Give an example that achieves this bound.

Solution 5. (a) Since X, Y, Z are pairwise independent fair flips, $H(X) = H(Y) = H(Z) = 1$, and $H(Y|X) = H(Y)$. Therefore, using the chain rule for entropy, $H(X, Y) = H(X) + H(Y|X) = H(X) + H(Y) = 2$.

(b) Using the chain rule for entropy, we can write: $H(X, Y, Z) = H(X, Y) + H(Z|X, Y) \geq H(X, Y) = 2$, which holds since (conditional) entropy is non-negative, thus $H(Z|X, Y) \geq 0$.

(c) Let $Z = X + Y \pmod{2}$, then $H(Z|X, Y) = 0$ and $H(X, Y, Z) = H(X, Y)$.

Problem 6: Entropy and Geometry

Suppose X, Y and Z are random variables.

- Show that $H(X) + H(Y) + H(Z) \geq \frac{1}{2}[H(X, Y) + H(Y, Z) + H(Z, X)]$.
- Show that $H(X, Y) + H(Y, Z) \geq H(X, Y, Z) + H(Y)$.
- Show that

$$2[H(X, Y) + H(Y, Z) + H(Z, X)] \geq 3H(X, Y, Z) + H(X) + H(Y) + H(Z).$$

- Show that $H(X, Y) + H(Y, Z) + H(Z, X) \geq 2H(X, Y, Z)$.
- Suppose n points in three dimensions are arranged so that their projections to the xy , yz and zx planes give n_{xy} , n_{yz} and n_{zx} points. Clearly $n_{xy} \leq n$, $n_{yz} \leq n$, $n_{zx} \leq n$. Use part (d) show that

$$n_{xy}n_{yz}n_{zx} \geq n^2.$$

Solution 6. (a) By the sub-additivity of Entropy we know that

$$\begin{aligned} H(X, Y) &\leq H(X) + H(Y) \\ H(Y, Z) &\leq H(Y) + H(Z) \\ H(X, Z) &\leq H(X) + H(Z). \end{aligned}$$

Adding the three inequalities together we retrieve:

$$H(X) + H(Y) + H(Z) \geq \frac{1}{2} (H(X, Y) + H(Y, Z) + H(Z, X)).$$

(b) It is easier to show

$$H(X, Y) + H(Y, Z) - (H(X, Y, Z) + H(Y)) \geq 0.$$

Indeed we have that:

$$H(X|Y) - H(X|Y, Z) = I(X; Z|Y) \geq 0.$$

(c) Applying (b), but inverting the roles of X, Y, Z we get:

$$\begin{aligned} H(X, Y) + H(Y, Z) &\geq H(X, Y, Z) + H(Y) \\ H(Y, Z) + H(Z, X) &\geq H(Y, Z, X) + H(Z) \\ H(Y, X) + H(X, Z) &\geq H(Y, X, Z) + H(X). \end{aligned}$$

Adding the three inequalities together gives us (c).

(d) By sub-additivity again, we have that:

$$H(X, Y, Z) \leq H(X) + H(Y) + H(Z). \quad (30)$$

Using (30) in (c) we retrieve

$$\begin{aligned} 2[H(X, Y) + H(Y, Z) + H(X, Z)] &\geq 3H(X, Y, Z) + H(X) + H(Y) + H(Z) \\ &\geq 3H(X, Y, Z) + H(X, Y, Z) \\ &= 4H(X, Y, Z). \end{aligned}$$

(e) Let $\{(x_i, y_i, z_i) : i = 1, \dots, n\}$ be our set of points. Suppose that X, Y, Z are random variables representing the components of the n points with respect to the x, y, z axes. Furthermore, suppose that three random variables are such that $\Pr((X, Y, Z) = (x_i, y_i, z_i)) = 1/n$ for every $1 \leq i \leq n$. This implies that

$$H(X, Y, Z) = \log n. \quad (31)$$

Consequently the random couples $(X, Y), (X, Z), (Y, Z)$ represent the projections of the points respectively, on the xy, xz and yz axes. We can thus say that

$$H(X, Y) \leq \log n_{xy} \quad (32)$$

$$H(X, Z) \leq \log n_{xz} \quad (33)$$

$$H(Y, Z) \leq \log n_{yz}. \quad (34)$$

Using (31),(32),(33),(34) in (d) we retrieve the following:

$$\log(n_{xy}n_{xz}n_{yz}) \geq H(X, Y) + H(Y, Z) + H(X, Z) \geq 2H(X, Y, Z) = 2\log n.$$

Which is equivalent to:

$$(n_{xy}n_{xz}n_{yz}) \geq n^2.$$

Problem 7: Entropy and combinatorics

Let $n \geq 1$ and fix some $0 \leq k \leq n$. Let $p = \frac{k}{n}$ and let $T_p^n \subset \{0,1\}^n$ be the set of all binary sequences with exactly np ones (assume that np is an integer).

(a) Show that

$$\log |T_p^n| = nh(p) + O(\log_e n)$$

where $h(p) = -p \log_e p - (1-p) \log_e (1-p)$ is the binary entropy function.

Hint: Stirling's approximation states that for every $n \geq 1$,

$$e^{\frac{1}{12n+1}} \sqrt{2\pi n} \left(\frac{n}{e}\right)^n \leq n! \leq e^{\frac{1}{12n}} \sqrt{2\pi n} \left(\frac{n}{e}\right)^n$$

(b) Let $Q^n = \text{Bernoulli}(q)^n$ be the i.i.d. Bernoulli distribution on $\{0,1\}^n$. Show that

$$\log Q^n[T_p^n] = -nd(p\|q) + O(\log_e n)$$

where $d(p\|q) = p \log_e \frac{p}{q} + (1-p) \log_e \frac{1-p}{1-q}$ is the binary KL divergence.

Solution 7. (a) When $p = 0$ or 1 , we have $|T_p^n| = 1$, or equivalently $\log |T_p^n| = 0$, so the result holds trivially, since $h(p) = 0$ for $p = 0, 1$. For $p \neq 0, 1$, we have that $|T_p^n| = \binom{n}{np} = \frac{n!}{(np)!(n(1-p))!}$. Using Stirling's approximation on the three factorials we get

$$\begin{aligned} \frac{1}{\sqrt{2\pi np(1-p)}} p^{-np} (1-p)^{-n(1-p)} e^{\frac{1}{12n+1} - \frac{1}{12np} - \frac{1}{12n(1-p)}} &\leq |T_p^n| \\ &\leq \frac{1}{\sqrt{2\pi np(1-p)}} p^{-np} (1-p)^{-n(1-p)} e^{\frac{1}{12n} - \frac{1}{12np+1} - \frac{1}{12n(1-p)+1}}. \end{aligned}$$

By taking the log on each side, we get

$$\begin{aligned} nh(p) - \frac{1}{2} \log(2\pi np(1-p)) + \frac{1}{12n+1} - \frac{1}{12np} - \frac{1}{12n(1-p)} &\leq \log |T_p^n| \\ &\leq nh(p) - \frac{1}{2} \log(2\pi np(1-p)) + \frac{1}{12n} - \frac{1}{12np+1} - \frac{1}{12n(1-p)+1}. \end{aligned}$$

Since $\frac{1}{n} \leq p \leq \frac{n-1}{n}$ and the same holds for $1-p$, we can obtain the following (loose) bounds:

$$\begin{aligned} -\frac{1}{2} \log n + \frac{1}{2} \log(2\pi) &\leq \frac{1}{2} \log(2\pi np(1-p)) \leq \frac{1}{2} \log n + \frac{1}{2} \log(2\pi) \\ \frac{1}{12n+1} - \frac{1}{12np} - \frac{1}{12n(1-p)} &\geq -2 \\ \frac{1}{12n} - \frac{1}{12np+1} - \frac{1}{12n(1-p)+1} &\leq 1 \end{aligned}$$

so that we get

$$nh(p) - \frac{1}{2} \log n - \frac{1}{2} \log(2\pi) - 2 \leq \log |T_p^n| \leq nh(p) + \frac{1}{2} \log n - \frac{1}{2} \log(2\pi) + 1$$

i.e., $\log |T_p^n| = nh(p) + O(\log n)$.

(b) We have

$$Q^n[T_p^n] = \binom{n}{np} q^{np} (1-q)^{n(1-p)} = |T_p^n| q^{np} (1-q)^{n(1-p)}$$

and therefore

$$\begin{aligned}
\log Q^n[T_p^n] &= \log |T_p^n| + np \log q + n(1-p) \log(1-q) \\
&= nh(p) + np \log q + n(1-p) \log(1-q) + O(\log n) \\
&= -nd(p||q) + O(\log n)
\end{aligned}$$

where in the last step we used (a).

Problem 8: Axiomatic definition of entropy

Let $(p_1, p_2, \dots, p_m)$ be such that $p_i \geq 0$ for $i = 1, \dots, m$ and $\sum_i p_i = 1$. Let

$$H_m(p_1, \dots, p_m) = - \sum_{i=1}^m p_i \log p_i \quad (35)$$

be the entropy of $(p_1, p_2, \dots, p_m)$.

(a) (*Grouping property*) Prove that

$$H_m(p_1, p_2, p_3, \dots, p_m) = H_{m-1}(p_1 + p_2, p_3, \dots, p_m) + (p_1 + p_2) H_2\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right).$$

Also prove it for grouping p_i and p_j for any arbitrary pair of indices (i, j) . This property models the fact that the uncertainty in choosing among m objects should be equal to the uncertainty in first choosing a subgroup of the objects, and then choosing an object in the selected subgroup.

(b) Prove that if a sequence of functions F_m of probability vectors $(p_1, p_2, \dots, p_m)$, is such that for every $m \geq 2$,

1. $F_m(p_1, p_2, \dots, p_m)$ is continuous in the p_i 's,
2. $F_m(p_1, p_2, \dots, p_m)$ satisfies the grouping property (a),
3. $F_m(\frac{1}{m}, \dots, \frac{1}{m}) = \log m$

then F_m must be equal to the entropy (35) (under the usual convention $0 \log 0 = 0$).

Hint: Suppose that the p_i 's are rational, i.e., $p_i = \frac{n_i}{n}$ for some positive integers $\{n_i\}_{i=1, \dots, m}$. Show using (a) recursively that

$$F_n\left(\frac{1}{n}, \dots, \frac{1}{n}\right) = F_m\left(\frac{n_1}{n}, \dots, \frac{n_m}{n}\right) + \sum_i \frac{n_i}{n} F_{n_i}\left(\frac{1}{n_i}, \dots, \frac{1}{n_i}\right).$$

Solution 8. (a) Using (35), we can rewrite the right-hand side as

$$\begin{aligned}
&H(p_1 + p_2, p_3, \dots, p_m) + (p_1 + p_2) H\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right) \\
&= -(p_1 + p_2) \log(p_1 + p_2) - \sum_{i=3}^m p_i \log p_i + (p_1 + p_2) \left(-\frac{p_1}{p_1 + p_2} \log \frac{p_1}{p_1 + p_2} - \frac{p_2}{p_1 + p_2} \log \frac{p_2}{p_1 + p_2} \right) \\
&= -(p_1 + p_2) \log(p_1 + p_2) - \sum_{i=3}^m p_i \log p_i - p_1 \log p_1 - p_2 \log p_2 + (p_1 + p_2) \log(p_1 + p_2) \\
&= - \sum_{i=1}^m p_i \log p_i = H(p_1, p_2, p_3, \dots, p_m).
\end{aligned}$$

(b) It can be proved by induction that the grouping property holds for grouping an arbitrary number of elements. Hence, using it recursively on $F\left(\frac{1}{m}, \dots, \frac{1}{m}\right)$, we get

$$F\left(\frac{1}{m}, \dots, \frac{1}{m}\right) = F\left(\frac{m_1}{m}, \dots, \frac{m_k}{m}\right) + \sum_i \frac{m_i}{m} F\left(\frac{1}{m_i}, \dots, \frac{1}{m_i}\right).$$

Using property 3 on $F\left(\frac{1}{m}, \dots, \frac{1}{m}\right)$ and on each $F\left(\frac{1}{m_i}, \dots, \frac{1}{m_i}\right)$, we get

$$\log m = F\left(\frac{m_1}{m}, \dots, \frac{m_k}{m}\right) + \sum_i \frac{m_i}{m} \log m_i.$$

Rearranging the last equation gives

$$F\left(\frac{m_1}{m}, \dots, \frac{m_k}{m}\right) = -\sum_i \frac{m_i}{m} \log \frac{m_i}{m}.$$

This proves the result for every rational probability vector. By using the continuity of F (property 1), we can extend the result to any probability vector.

Problem 9: Conditional KL divergence

We saw in class that a *probability kernel* $P_{Y|X} : \mathcal{X} \rightarrow \mathcal{Y}$ is a matrix $P_{Y|X} = P_{Y|X}(y|x) : x \in \mathcal{X}, y \in \mathcal{Y}$ such that $P_{Y|X}(y|x) \geq 0$, and for each $x \in \mathcal{X}$, $\sum_y P_{Y|X}(y|x) = 1$. Let $P_X \in \Pi(\mathcal{X})$ be a probability distribution on $\mathcal{X}$. We define the *conditional KL divergence* between two probability kernels $P_{Y|X} : \mathcal{X} \rightarrow \mathcal{Y}$ and $Q_{Y|X} : \mathcal{X} \rightarrow \mathcal{Y}$ given P_X to be

$$D(P_{Y|X} \| Q_{Y|X} | P_X) \triangleq \sum_{x \in \mathcal{X}} P_X(x) D(P_{Y|X}(\cdot|x) \| Q_{Y|X}(\cdot|x))$$

where for every x , $D(P_{Y|X}(\cdot|x) \| Q_{Y|X}(\cdot|x))$ is the standard KL divergence between the two distributions $P_{Y|X}(\cdot|x)$ and $Q_{Y|X}(\cdot|x)$ over $\mathcal{Y}$.

(a) (*Chain rule of the KL divergence*) Show that

$$D(P_{X,Y} \| Q_{X,Y}) = D(P_X \| Q_X) + D(P_{Y|X} \| Q_{Y|X} | P_X)$$

where $P_{X,Y}$ and $Q_{X,Y}$ are two joint distributions on $\mathcal{X} \times \mathcal{Y}$ such that $P_{X,Y}(x, y) = P_X(x)P_{Y|X}(y|x)$ and $Q_{X,Y}(x, y) = Q_X(x)Q_{Y|X}(y|x)$.

(b) Using (a), show that

$$D(P_{Y|X} \| Q_{Y|X} | P_X) = D(P_{X,Y} \| Q_{X,Y})$$

where $P_{X,Y}(x, y) = P_X(x)P_{Y|X}(y|x)$ and $Q_{X,Y}(x, y) = P_X(x)Q_{Y|X}(y|x)$.

(c) (*Conditioning increases divergence*) Using (b) and the Data Processing Inequality seen in class, show that

$$D(P_Y \| Q_Y) \leq D(P_{Y|X} \| Q_{Y|X} | P_X)$$

where $P_Y(y) = \sum_{x \in \mathcal{X}} P_X(x)P_{Y|X}(y|x)$ and $Q_Y(y) = \sum_{x \in \mathcal{X}} P_X(x)Q_{Y|X}(y|x)$.

Solution 9. (a)

$$\begin{aligned}
D(P_{XY} \| Q_{XY}) &= \sum_{x,y} P_{XY}(x,y) \log \frac{P_{XY}(x,y)}{Q_{XY}(x,y)} \\
&= \sum_{x,y} P_X(x) P_{Y|X}(y|x) \log \frac{P_X(x) P_{Y|X}(y|x)}{Q_X(x) Q_{Y|X}(y|x)} \\
&= \sum_{x,y} P_X(x) P_{Y|X}(y|x) \log \frac{P_X(x)}{Q_X(x)} + \sum_{x,y} P_X(x) P_{Y|X}(y|x) \log \frac{P_{Y|X}(y|x)}{Q_{Y|X}(y|x)} \\
&= D(P_X \| Q_X) + \sum_x P_X(x) D(P_{Y|X}(\cdot|x) \| Q_{Y|X}(\cdot|x)) = D(P_X \| Q_X) + D(P_{Y|X} \| Q_{Y|X} | P_X).
\end{aligned}$$

(b)

$$D(P_{XY} \| Q_{XY}) = D(P_X \| P_X) + D(P_{Y|X} \| Q_{Y|X} | P_X) = D(P_{Y|X} \| Q_{Y|X} | P_X).$$

(c) Define the kernel

$$W(\tilde{y}|x, y) = \begin{cases} 1, & \text{if } \tilde{y} = y, \\ 0, & \text{otherwise.} \end{cases}$$

Then we have $P_{\tilde{Y}}(\tilde{y}) = \sum_{x,y} P_{XY}(x,y) W(\tilde{y}|x, y) = P_Y(\tilde{y})$ and $Q_{\tilde{Y}}(\tilde{y}) = \sum_{x,y} Q_{XY}(x,y) W(\tilde{y}|x, y) = Q_Y(\tilde{y})$. Hence, we have

$$D(P_{Y|X} \| Q_{Y|X} | P_X) = D(P_{XY} \| Q_{XY}) \geq D(P_{\tilde{Y}} \| Q_{\tilde{Y}}) = D(P_Y \| Q_Y).$$

where the equality follows from part (b) and the inequality follows from DPI.