Problem Set 2 **(Graded)** —*Due Tuesday, October 10, before class starts*

For the Exercise Sessions on September 26 and Oct 3

Last name	First name	SCIPER Nr	Points

Problem 1: Axiomatic definition of entropy

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

$$H(p_1, \ldots, p_m) = -\sum_i p_i \log p_i \tag{1}$$

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

(a) *(Grouping property)* Prove that

$$H(p_1, p_2, p_3, \ldots, p_m) = H(p_1 + p_2, p_3, \ldots, p_m) + (p_1 + p_2) H\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right).$$

The above 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 function F of probability vectors $(p_1, p_2, \ldots, p_m)$, $m \geq 2$, is such that

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

then F must be equal to the entropy (1).

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

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

Solution 1. *(a)* Using (1), we can rewrite the right-hand side as

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

(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 2: Entropy and Geometry

Suppose X, Y and Z are random variables.

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

(b) Show that $H(X,Y) + H(Y,Z) \geq H(X,Y,Z) + H(Y)$.

(c) Show that

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

(d) Show that $H(X,Y) + H(Y,Z) + H(Z,X) \geq 2H(X,Y,Z)$.

(e) Suppose n points in three dimensions are arranged so that their 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 2. *(a)* By the sub-addivitity of Entropy we know that

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

Adding the three inequalities together we retrieve:

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

(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-addivity again, we have that:

$$H(X,Y,Z) \leq H(X) + H(Y) + H(Z). \tag{2}$$

Using (2) in *(c)* we retrieve

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

(d) 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. Furthemore, 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

$$H(X,Y,Z) = \log n. \tag{3}$$

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

$$H(X,Y) \leq \log n_{xy} \tag{4}$$
$$H(X,Z) \leq \log n_{xz} \tag{5}$$
$$H(Y,Z) \leq \log n_{yz}. \tag{6}$$

Using (3),(4),(5),(6) 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 3: Conditional KL divergence

We saw in class that a *probability kernel* $P_{Y|X} : \mathcal{X} \to \mathcal{Y}$ 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 $P_{Y|X} : \mathcal{X} \to \mathcal{Y}$ and $Q_{Y|X} : \mathcal{X} \to \mathcal{Y}$ 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

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

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 3. *(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, by the DPI we have

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

Problem 4: 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 $E_{P_XP_Y}[e^{f(X,Y)}]$ is finite,

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

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

Solution 4. *(a)*

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

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

(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 5: f-divergences

Suppose f is a convex function defined on $(0,\infty)$ with $f(1)=0$. Define the f-divergence of a distribution P from a distribution Q as

$$D_f(P\|Q) \triangleq \sum_x Q(x) f(P(x)/Q(x)).$$

In the sum above we take $f(0) := \lim_{t\to 0} f(t)$, $0f(0/0) := 0$, and $0f(a/0) := \lim_{t\to 0} tf(a/t) = a\lim_{t\to 0} tf(1/t)$.

(a) Show that the following basic properties hold:

1. $D_{f_1+f_2}(P\|Q) = D_{f_1}(P\|Q) + D_{f_2}(P\|Q)$
2. $D_f(P\|P) = 0$
3. $D_f(P\|Q) \ge 0$

(b) *(Monotonicity)* Show that $D_f(P_{XY}\|Q_{XY}) \ge D_f(P_X\|Q_X)$.

(c) *(Data processing inequality)* Show that for any probability kernel $W(y|x)$ from $\mathcal{X}$ to $\mathcal{Y}$, and any two distributions P_X and Q_X on $\mathcal{X}$

$$D_f(P_X\|Q_X) \ge D_f(P_Y\|Q_Y)$$

where P_Y and Q_Y are probability distributions on $\mathcal{Y}$ given by $P_Y(y) = \sum_x P_X(x)W(y|x)$ and $Q_Y(y) = \sum_x Q_X(x)W(y|x)$.

(d) Show that if f is strictly convex in 1, then $D_f(P\|Q) = 0$ if and only if $P = Q$.

Solution 5. *(a)*

1.

$$\begin{aligned}D_{f_1+f_2}(P\|Q) &= \sum_x Q(x)\left[f_1(P(x)/Q(x)) + f_2(P(x)/Q(x))\right] \\ &= \sum_x Q(x) f_1(P(x)/Q(x)) + \sum_x Q(x) f_2(P(x)/Q(x)) \\ &= D_{f_1}(P\|Q) + D_{f_2}(P\|Q).\end{aligned}$$

2. $D_f(P\|P) = \sum_x P(x) f(P(x)/P(x)) = \sum_x P(x) f(1) = 0.$

3. $D_f(P\|Q) = \sum_x Q(x) f(P(x)/Q(x)) \geq f\left(\sum_x Q(x)\frac{P(x)}{Q(x)}\right) = f\left(\sum_x P(x)\right) = f(1) = 0$ where we used Jensen's inequality since f is convex.

(b)

$$\begin{aligned}
D_f(P_{XY}\|Q_{XY}) &= \sum_{x,y} Q_{XY}(x,y) f\left(\frac{P_{XY}(x,y)}{Q_{XY}(x,y)}\right) \\
&= \sum_x Q_X(x) \sum_y Q_{Y|X}(y|x) f\left(\frac{P_{XY}(x,y)}{Q_{XY}(x,y)}\right) \\
&\geq \sum_x Q_X(x) f\left(\sum_y Q_{Y|X}(y|x)\frac{P_{XY}(x,y)}{Q_{XY}(x,y)}\right) \\
&= \sum_x Q_X(x) f\left(\frac{\sum_y P_{XY}(x,y)}{Q_X(x)}\right) \\
&= \sum_x Q_X(x) f\left(\frac{P_X(x)}{Q_X(x)}\right) = D_f(P\|Q)
\end{aligned}$$

where the inequality is again due to Jensen.

(c) From (b) we have $D_f(P_{XY}\|Q_{XY}) \geq D_f(P_Y\|Q_Y)$. But we also have

$$\begin{aligned}
D_f(P_{XY}\|Q_{XY}) &= \sum_{x,y} Q_X(x) W(y|x) f\left(\frac{P_X(x)W(y|x)}{Q_X(x)W(y|x)}\right) \\
&= \sum_x Q_X(x)\left(\sum_y W(y|x)\right) f\left(\frac{P_X(x)}{Q_X(x)}\right) \\
&= D_f(P_X\|Q_X)
\end{aligned}$$

that is, $D_f(P_X\|Q_X) \geq D_f(P_Y\|Q_Y)$.

(d) Since f is strictly convex in 1, for every $s,t > 0$ and $0 < \alpha < 1$ such that $\alpha s + (1-\alpha)t = 1$, we have $\alpha f(s) + (1-\alpha) f(t) > f(1) = 0$. Suppose by contradiction that $P \neq Q$ and $D_f(P\|Q) = 0$. Then there exists $\tilde{x}$ such that $P(\tilde{x}) \neq Q(\tilde{x})$. Define the random variable $Y = 1_{\{X=\tilde{x}\}}$, and let $p \triangleq P(\tilde{x})$ and $q \triangleq Q(\tilde{x})$. Using (c) we get that $0 \leq D_f(P_Y\|Q_Y) = D_f(p\|q) \leq D_f(P\|Q) = 0$, i.e., $D_f(p\|q) = qf\left(\frac{p}{q}\right) + (1-q) f\left(\frac{1-p}{1-q}\right) = 0$. But this contradicts the fact that f is strictly convex in 1, since if you set $s = \frac{p}{q}$, $t = \frac{1-p}{1-q}$ and $\alpha = q$, the last equation can be rewritten as $\alpha f(s) + (1-\alpha) f(t) = 0$, a contradiction.

Problem 6: 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.

(a) Show that

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

where $h(p) = -p\log p - (1-p)\log(1-p)$ 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

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

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

Solution 6. *(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:

$$-\frac{1}{2}\log n + \frac{1}{2}\log(2\pi) \leq \frac{1}{2}\log(2\pi np(1-p)) \leq \frac{1}{2}\log n + \frac{1}{2}\log(2\pi)$$

$$\frac{1}{12n+1} - \frac{1}{12np} - \frac{1}{12n(1-p)} \geq -2$$

$$\frac{1}{12n} - \frac{1}{12np+1} - \frac{1}{12n(1-p)+1} \leq 1$$

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).