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*

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

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

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

(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

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

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

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

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

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 $E_{P_X P_Y}[e^{f(X,Y)}]$ 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 $E_{P_X P_Y}[e^{f(X,Y)}]$ 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.

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. $D_f(P\|Q) \geq 0$

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

(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 $\mathcal{X}$

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

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 $P = Q$.

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.