Foundations of Data Science
Gastpar & Urbanke

Ecole Polytechnique Fédérale, Lausanne: Fall 2026
September 15, 2026

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

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

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.

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 $E_{P_X P_Y}[e^{f(X,Y)}]$ 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 4: Geometrical interpretation of mutual information

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 $x \in \mathcal{X}$, $D(P_{Y|X}(\cdot|x) \| Q_{Y|X}(\cdot|x))$ 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.

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.

(a) What is $H(X, Y)$?

(b) Give a lower bound to the value of $H(X,Y,Z)$.

(c) Give an example that achieves this bound.

Problem 6: Entropy and Geometry

Suppose X, Y and Z are random variables.

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

(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 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.

Problem 8: 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

$$H_m(p_1, \ldots, p_m) = -\sum_{i=1}^{m} p_i \log p_i \tag{1}$$

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

(a) *(Grouping property)* Prove that

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

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 $(p_1, p_2, \ldots, p_m)$, is such that for every $m \geq 2$,

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

then F_m must be equal to the entropy (1) (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 $\{n_i\}_{i=1,\ldots,m}$. Show using (a) recursively that

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

Problem 9: 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 $Q_Y(y) = \sum_{x \in \mathcal{X}} P_X(x) Q_{Y|X}(y|x)$.