Foundations of Data Science
Urbanke

Ecole Polytechnique Fédérale, Lausanne: Fall 2022
September 30, 2022

Problem Set 2 —*Due Friday, October 14, before class starts*

For the Exercise Sessions on September 30 and Oct 7

Last name	First name	SCIPER Nr	Points

Problem 1: Entropy and pairwise independence

Suppose X, Y, Z are pairwise independent fair flips, i.e., $I(X;Y) = I(Y;Z) = I(Z;X) = 0$.

(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 2: Divergence and L_1

Suppose p and q are two probability mass functions on a finite set $\mathcal{U}$. (I.e., for all $u \in \mathcal{U}$, $p(u) \geq 0$ and $\sum_{u\in\mathcal{U}} p(u) = 1$; similarly for q.)

(a) Show that the L_1 distance $\|p - q\|_1 := \sum_{u\in\mathcal{U}} |p(u) - q(u)|$ between p and q satisfies

$$\|p - q\|_1 = 2 \max_{\mathcal{S}:\mathcal{S}\subset\mathcal{U}} p(\mathcal{S}) - q(\mathcal{S})$$

with $p(\mathcal{S}) = \sum_{u\in\mathcal{S}} p(u)$ (and similarly for q), and the maximum is taken over all subsets $\mathcal{S}$ of $\mathcal{U}$.

For α and β in $[0,1]$, define the function $d_2(\alpha\|\beta) := \alpha \log \frac{\alpha}{\beta} + (1-\alpha)\log\frac{1-\alpha}{1-\beta}$. Note that $d_2(\alpha\|\beta)$ is the divergence of the distribution $(\alpha, 1-\alpha)$ from the distribution $(\beta, 1-\beta)$.

(b) Show that the first and second derivatives of d_2 with respect to its first argument α satisfy $d_2'(\beta\|\beta) = 0$ and $d_2''(\alpha\|\beta) = \frac{\log e}{\alpha(1-\alpha)} \geq 4\log e$.

(c) By Taylor's theorem conclude that

$$d_2(\alpha\|\beta) \geq 2(\log e)(\alpha - \beta)^2.$$

(d) Show that for any $\mathcal{S} \subset \mathcal{U}$

$$D(p\|q) \geq d_2(p(\mathcal{S})\|q(\mathcal{S}))$$

[Hint: use the data processing theorem for divergence.]

(e) Combine (a), (c) and (d) to conclude that

$$D(p\|q) \geq \tfrac{\log e}{2}\|p - q\|_1^2.$$

(f) Show, by example, that $D(p\|q)$ can be $+\infty$ even when $\|p-q\|_1$ is arbitrarily small. [Hint: considering $\mathcal{U}=\{0,1\}$ is sufficient.] Consequently, there is no generally valid inequality that upper bounds $D(p\|q)$ in terms of $\|p-q\|_1$.

Problem 3: Generating fair coin flips from rolling the dice

Suppose $X_1, X_2, \ldots$ are the outcomes of rolling a possibly loaded die multiple times. The outcomes are assumed to be iid. Let $\mathbb{P}(X_i = m) = p_m$, for $m = 1, 2, \ldots, 6$, with p_m unknown (but non-negative and summing to one, clearly). By processing this sequence we would like to obtain a sequence $Z_1, Z_2, \ldots$ of *fair* coin flips.

Consider the following method: We process the X sequence in successive pairs, (X_1X_2), (X_3X_4), (X_5X_6), mapping $(3,4)$ to 0, $(4,3)$ to 1, and all the other outcomes to the empty string λ. After processing X_1, X_2, we will obtain either nothing, or a bit Z_1.

(a) Show that, if a bit is obtained, it is fair, i.e., $\mathbb{P}(Z_1 = 0|Z_1 \neq \lambda) = \mathbb{P}(Z_1 = 1|Z_1 \neq \lambda) = 1/2$.

In general we can process the X sequence in successive n-tuples via a function $f : \{1,2,3,4,5,6\}^n \to \{0,1\}^*$ where $\{0,1\}^*$ denotes the set of all finite length binary sequences (including the empty string λ). [The case in (a) is the function where $f(3,4) = 0$, $f(4,3) = 1$, and $f(j,m) = \lambda$ for all other choices of j and m.] The function f is chosen such that $(Z_1, \ldots, Z_K) = f(X_1, \ldots, X_n)$ are i.i.d., and fair (here K may depend on $(X_1, \ldots, X_n)$).

(b) Letting $H(X)$ denote the entropy of the (unknown) distribution $(p_1, p_2, \ldots, p_6)$, prove the following chain of (in)equalities.

$$\begin{aligned} nH(X) &= H(X_1, \ldots, X_n) \\ &\geq H(Z_1, \ldots, Z_K, K) \\ &= H(K) + H(Z_1 \ldots, Z_K|K) \\ &= H(K) + \mathbb{E}[K] \\ &\geq \mathbb{E}[K]. \end{aligned}$$

Consequently, on the average no more than $nH(X)$ fair bits can be obtained from $(X_1, \ldots, X_n)$.

(c) Describe how you would find a good f (with high $\mathbb{E}[K]$) for $n = 4$ which would work for any distribution $(p_1, p_2, ..., p_6)$.

Advanced Problems

Problem 4: Extremal characterization for Rényi entropy

Given $s \geq 0$, and a random variable U taking values in $\mathcal{U}$, with probabilitis $p(u)$, consider the distribution $p_s(u) = p(u)^s / Z(s)$ with $Z(s) = \sum_u p(u)^s$.

(a) Show that for any distribution q on $\mathcal{U}$,

$$(1-s)H(q) - sD(q\|p) = -D(q\|p_s) + \log Z(s).$$

(b) Given s and p, conclude that the left hand side above is maximized by the choice by $q = p_s$ with the value $\log Z(s)$,

The quantity

$$H_s(p) := \frac{1}{1-s} \log Z(s) = \frac{1}{1-s} \log \sum_u p(u)^s$$

is known as the *Rényi entropy of order s of the random variable U*. When convenient, we will also write $H_s(U)$ instead of $H_s(p)$.

(c) Show that if U and V are independent random variables

$$H_s(UV) := H_s(U) + H_s(V).$$

[Here UV denotes the pair formed by the two random variables — not their product. E.g., if $\mathcal{U} = \{0, 1\}$ and $\mathcal{V} = \{a, b\}$, UV takes values in $\{0a, 0b, 1a, 1b\}$.]

Problem 5: KL and its Fenchel-Legendre dual

Consider the Kullback-Leibler divergence $D(Q||P)$ as a function of Q, for fixed P.

(a) Show that its convex conjugate (sometimes also called Fenchel-Legendre dual) is the logarithm of the moment-generating function of P. *Hint:* To keep arguments simple, assume that P is a finite-dimensional probability mass function, thus $P \in \mathbb{R}^n$, and that $P(x) > 0$ for all x. Recall that the convex conjugate is the function $f^*(Q^*) = \sup_Q \langle Q^*, Q \rangle - D(Q||P)$, where $Q^* \in \mathbb{R}^n$.

(b) Fix P to be a normal distribution of mean zero. Let Q be arbitrary but with the same second moment as P. Show that in this case, $D(Q||P) = h(P) - h(Q)$, that is, the difference of the differential entropy of the normal distribution and the differential entropy of Q.

Problem 6: Moments and Rényi

Suppose G is an integer valued random variable taking values in the set $\{1, \ldots, K\}$. Let $p_i = \Pr(G = i)$. We will derive bounds on the moments of G, the ρ-th moment of G being $\mathbb{E}[G^\rho]$.

1. Show that for any distribution q on $\{1, \ldots, K\}$, and any ρ

$$\mathbb{E}[G^\rho] = \sum_i q_i \exp\Big[\log \frac{p_i i^\rho}{q_i}\Big].$$

(Here and below exp and log are taken to same base.)

2. Show that

$$\mathbb{E}[G^\rho] \geq \exp\Big[-D(q\|p) + \rho \sum_i q_i \log i\Big].$$

[*Hint:* use Jensen's inequality on Part 1.]

3. Show that
$$\sum_i q_i \log i = H(q) - \sum_i q_i \log \frac{1}{iq_i} \geq H(q) - \log \sum_{i=1}^{K} 1/i.$$
[*Hint:* use Jensen's inequality.]

4. Using Part 2, Part 3, and the fact that $\sum_{i=1}^{K} 1/i \leq 1 + \ln K$, show that, for $\rho \geq 0$,
$$\mathbb{E}[G^\rho] \geq (1 + \ln K)^{-\rho} \exp[\rho H(q) - D(q\|p)]$$

5. Suppose that $U_1, \ldots, U_n$ are i.i.d., each with distribution p. Suppose we try to determine the value of $X = (U_1, \ldots, U_n)$ by asking a sequence of questions, each of the type 'Is $X = x$?' until we are answered 'yes'. Let G_n be the number of questions we ask.

   Show that, for $\rho \geq 0$,
$$\liminf_n \frac{1}{n\rho} \log \mathbb{E}[G_n^\rho] \geq H_{1/(1+\rho)}(p)$$
   where $H_s(p) = \frac{1}{1-s} \log \sum_u p(u)^s$ is the Rényi entropy of the distribution p.

   [*Hint:* recall from Homework 2 Problem 6 that $\rho H_{1/(1+\rho)}(p) = \max_q \rho H(q) - D(q\|p)$, and that the Rényi entropy of a collection of independent random variables is the sum of their Rényi entropies.]

Problem 7: Other 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
$$D_f(p\|q) := \sum_u q(u) f(p(u)/q(u)).$$
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 for any non-negative a_1, a_2, b_1, b_2 and with $A = a_1 + a_2$, $B = b_1 + b_2$,
$$b_1 f(a_1/b_1) + b_2 f(a_2/b_2) \geq Bf(A/B);$$
and that in general, for any non-negative $a_1, \ldots, a_k$, $b_1, \ldots, b_k$, and $A = \sum_i a_i$, $B = \sum_i b_i$, we have
$$\sum_i b_i f(a_i/b_i) \geq Bf(A/B).$$
[Hint: since f is convex, for any $\lambda \in [0, 1]$ and any $x_1, x_2 > 0$ $\lambda f(x_1) + (1 - \lambda) f(x_2) \geq f(\lambda x_1 + (1 - \lambda) x_2)$; consider $\lambda = b_1/B$.]

(b) Show that $D_f(p\|q) \geq 0$.

(c) Show that D_f satisfies the data processing inequality: for any transition probability kernel $W(v|u)$ from $\mathcal{U}$ to $\mathcal{V}$, and any two distributions p and q on $\mathcal{U}$
$$D_f(p\|q) \geq D_f(\tilde{p}\|\tilde{q})$$
where $\tilde{p}$ and $\tilde{q}$ are probability distributions on $\mathcal{V}$ defined via $\tilde{p}(v) := \sum_u W(v|u)p(u)$, and $\tilde{q}(v) := \sum_u W(v|u)q(u)$,

(d) Show that each of the following are f-divergences.

   i. $D(p\|q) := \sum_u p(u) \log(p(u)/q(u))$. [Warning: $\log$ is not the right choice for f.]
   ii. $R(p\|q) := D(q\|p)$.
   iii. $1 - \sum_u \sqrt{p(u)q(u)}$
   iv. $\|p - q\|_1$.
   v. $\sum_u (p(u) - q(u))^2/q(u)$