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.

Solution 1. *(a)* Since X, Y, Z are pairwise independent fair flips, $H(X) = H(Y) = H(Z) = 1$. $H(X,Y) = H(X) + H(Y|X) = H(X) + H(Y) - I(X;Y) = 2$.

(b) $H(X,Y,Z) = H(X,Y) + H(Z|X,Y) \geq H(X,Y) = 2$

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

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 $\|p-q\|_1$.

Solution 2. *(a)* For any set $\mathcal{S}$, we have
$$p(\mathcal{S}) - q(\mathcal{S}) = \sum_{u\in\mathcal{S}} p(u) - q(u) \leq \sum_{u\in\mathcal{S}} |p(u)-q(u)|. \tag{1}$$

Similarly for the compliment set of $\mathcal{S}$, we also have
$$q(\mathcal{S}^c) - p(\mathcal{S}^c) = \sum_{u\in\mathcal{S}^c} q(u) - p(u) \leq \sum_{u\in\mathcal{S}^c} |p(u)-q(u)|. \tag{2}$$

Note that $p(\mathcal{S}) + p(\mathcal{S}^c) = q(\mathcal{S}) + q(\mathcal{S}^c) = 1$. Thus $q(\mathcal{S}^c) - p(\mathcal{S}^c) = p(\mathcal{S}) - q(\mathcal{S})$. Therefore, we have
$$2(p(\mathcal{S}) - q(\mathcal{S})) \leq \sum_{u\in\mathcal{S}} |p(u)-q(u)| + \sum_{u\in\mathcal{S}^c} |p(u)-q(u)| = \sum_{u\in\mathcal{U}} |p(u)-q(u)| = \|p-q\|_1 \tag{3}$$

For the choice $\mathcal{S} = \{u : p(u) > q(u)\}$, we have
$$p(\mathcal{S}) - q(\mathcal{S}) = \sum_{u\in\mathcal{S}} p(u) - q(u) = \sum_{u\in\mathcal{S}} |p(u)-q(u)| \tag{4}$$

$$q(\mathcal{S}^c) - p(\mathcal{S}^c) = \sum_{u\in\mathcal{S}^c} q(u) - p(u) = \sum_{u\in\mathcal{S}^c} |p(u)-q(u)| \tag{5}$$

So, for this $\mathcal{S}$, we have $2(p(\mathcal{S}) - q(\mathcal{S})) = \|p-q\|_1$.

(b): Since $d_2(\alpha||\beta) = \alpha \log\frac{\alpha}{\beta} + (1-\alpha)\log\frac{1-\alpha}{1-\beta}$,
$$d_2'(\alpha||\beta) = \frac{\partial d_2(\alpha||\beta)}{\partial\alpha} = \log\frac{\alpha}{\beta} + \log e - \log\frac{1-\alpha}{1-\beta} - \log e = \log\frac{\alpha(1-\beta)}{\beta(1-\alpha)} \tag{6}$$

Therefore, we have $d_2'(\beta||\beta) = 0$.
$$d_2''(\alpha||\beta) = \frac{\log e}{\alpha(1-\alpha)} \geq 4\log e \tag{7}$$

where equality achieves when $\alpha = 1/2$.

(c): Using Taylor's theorem together with the Lagrange form of the remainder we see that for any f for which f' is continuous,
$$f(\alpha) = f(\beta) + (\alpha-\beta)f'(\beta) + (1/2)(\alpha-\beta)^2 f''(x_i) \tag{8}$$

where x_i is a value between α and β. With $f(\alpha) = d_2(\alpha\|\beta)$, we thus have
$$d_2(\alpha\|\beta) = 0 + 0 + (1/2)(\alpha-\beta)^2 f''(x_i) \geq 2\log(e)(\alpha-\beta)^2 \tag{9}$$

(d) Consider a deterministic channel with binary output

$$V = \begin{cases} 1, & \text{if } V \in \mathcal{S} \\ 0, & \text{if } V \notin \mathcal{S} \end{cases} \tag{10}$$

Thus,

$$\begin{aligned} d_2(p(\mathcal{S})\|q(\mathcal{S})) &= p(\mathcal{S})\log\frac{p(\mathcal{S})}{q(\mathcal{S})} + (1-p(\mathcal{S}))\log\frac{1-p(\mathcal{S})}{1-q(\mathcal{S})} & (11)\\ &= p(V=1)\log\frac{p(V=1)}{q(V=1)} + p(V=0)\log\frac{p(V=0)}{q(V=0)} & (12)\\ &= D(p_V\|q_V) & (13) \end{aligned}$$

By data processing theorem for divergence, $D(p\|q) \geq D(p_V\|q_V)$

(e) Combine (a),(c) and (d) and choosing $\mathcal{S} = \{u : p(u) > q(u)\}$, we have $\forall \mathcal{S}$

$$D(p\|q) \geq d_2(p(\mathcal{S})\|q(\mathcal{S})) \geq 2(\log e)(p(\mathcal{S}) - q(\mathcal{S}))^2 = \frac{\log e}{2}\|p-q\|_1^2 \tag{14}$$

(f) Let p be Bernoulli distribution with probability ϵ to be 1 and q is 0 with probability 1. Then

$$D(p\|q) = p(1)\log\frac{p(1)}{q(1)} + p(0)\log\frac{p(0)}{q(0)} = +\infty \tag{15}$$

But $\|p-q\|_1 = 2\epsilon$.

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 $(p_1, p_2, ..., p_6)$.

Solution 3. *(a)* $P(Z_1 = 0|Z_1 \neq \lambda) = P(Z_1 = 0, Z_1 \neq \lambda)/P(Z_1 \neq \lambda) = P(Z_1 = 0)/P(Z_1 \neq \lambda)$. Similarly, $P(Z_1 = 1|Z_1 \neq \lambda) = P(Z_1 = 1)/P(Z_1 \neq \lambda)$. Let us now show that $P(Z_1 = 0) = P(Z_1 = 1)$ and this will complete the proof. Note that $P(Z_1 = 1) = P(X_1 = 3, X_2 = 4) = P(X_1 = 3)P(X_2 = 4) = p_3 p_4$ and $P(Z_1 = 0) = P(X_1 = 4, X_2 = 3) = P(X_1 = 4)P(X_2 = 3) = p_4 p_3$. Therefore $P(Z_1 = 1) = P(Z_1 = 0)$.

(b)

$$
\begin{aligned}
nH(X) &= nH(X_i) && (16)\\
&= H(X_1, \ldots, X_n) \quad \text{[Independence of } X_i\text{]} && (17)\\
&\geq H(f(X_1, \ldots, X_n)) \quad \text{[Data Processing Inequality]} && (18)\\
&= H(Z_1, \ldots, Z_K, K) && (19)\\
&= H(K) + H(Z_1, \ldots, Z_K|K) \quad \text{[Chain Rule]} && (20)\\
&= H(K) + \sum_k p(K = k) H(Z_1, \ldots, Z_K|K = k) && (21)\\
&= H(K) + \sum_k p(K = k) k \quad [Z_1, \ldots, Z_k \text{ are i.i.d and fair when } K = k] && (22)\\
&= H(K) + \mathbb{E}[K] && (23)\\
&\geq \mathbb{E}[K] \quad \text{[Non-negativity of entropy]} && (24)
\end{aligned}
$$

(c)
We have in total 6^4 many possible outcomes. We can only produce fair bits, regardless of the distribution, if we have permutations of the same sequence. e.g., $1555 \to 00, 5155 \to 01, 5515 \to 10, 5551 \to 11$. Let us do the counting. A sequence can have $1, 2, 3$ or 4 kinds of different symbols. An example to a sequence of 3 different symbols is 1232.

1: We cannot produce bits with 1 kind of different symbols because you cannot permute the sequence and get another sequence. Therefore we map sequences of kind $aaaa$ to the null string λ.
2: For 2 different symbols it will be either 3 of the same kind and 1 of another kind which gives 4 different permutations or 2 of the same kind and 2 of another kind, which gives 6 different permutations. From the 4 different permutations of a " 3 by 1 " (aaab) sequence we can generate 2 fair bits, because there are 4 permutations. From the the first 4 of the 6 different permutations of a " 2 by 2 " sequence (aabb) we can generate 2 fair bits, and from the remaining 2 permutations we can generate 1 fair bit.
3: For 3 different symbols it has to be 2 of the same symbol, 1 of another symbol and 1 of another symbol $(aabc)$. There are $4!/2! = 12$ different ways to permute these sequence of type $aabc$. From the first 8 we can generate 3 bits, and from the remaining 4 we can generate 2 bits.
4: There are $4! = 24$ ways to permute a sequence of kind (a, b, c, d). From the first 16 we can generate 4 bits, and from the remaining 8 we can generate 3 bits.

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 $\{0a, 0b, 1a, 1b\}$.]

Solution 4. *(a)* We start from the left hand side of the equation:

$$\begin{aligned}
(1-s)H(q) - sD(q\|p) &= (1-s)\sum_u q(u)\log\frac{1}{q(u)} - s\sum_u q(u)\log\frac{q(u)}{p(u)} && (25)\\
&= \sum_u q(u)\left((1-s)\log\frac{1}{q(u)} - s\log\frac{q(u)}{p(u)}\right) && (26)\\
&= \sum_u q(u)\log\frac{p(u)^s}{q(u)} && (27)\\
&= \sum_u q(u)\log\frac{p_s(u)Z(s)}{q(u)} && (28)\\
&= \sum_u q(u)\log\frac{p_s(u)}{q(u)} + \sum_u q(u)\log Z(s) && (29)\\
&= -D(q\|p_s) + \log Z(s) && (30)
\end{aligned}$$

(b) We know that $D(q\|p_s) \geq 0$, where equality achieves for $q = p_s$. The left hand side of above equation is maximized when $q = p_s$ and has value $\log Z(s)$.

(c) Since U and V are independent random variables, we have $p(u,v) = p(u)p(v)$.

$$\begin{aligned} H_s(UV) &= \frac{1}{1-s} \log \sum_{u,v} p(u,v)^s && (31) \\ &= \frac{1}{1-s} \log(\sum_u p(u)^s \sum_v p(v)^s) && (32) \\ &= \frac{1}{1-s} \log \sum_u p(u)^s + \frac{1}{1-s} \log \sum_v p(v)^s && (33) \\ &= H_s(U) + H_s(V) && (34) \end{aligned}$$

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.

Solution 5. *(a)* The Lagrangian is

$$L(\lambda, Q) = \left(\sum_{x\in\mathcal{X}} Q^*(x)Q(x)\right) - \sum_{x\in\mathcal{X}} Q(x)\log\frac{Q(x)}{P(x)} - \lambda\left(\sum_{x\in\mathcal{X}} Q(x) - 1\right) \tag{35}$$

Taking the derivative with respect to $Q(x)$ gives

$$\frac{d}{dQ(x)} L(\lambda, Q) = Q^*(x) - \log\frac{Q(x)}{P(x)} - 1 - \lambda \tag{36}$$

Setting this to zero, we find

$$Q(x) = P(x)e^{Q^*(x)-(1+\lambda)}, \tag{37}$$

where we observe that $Q(x)$ is non-negative (which is good). Next, we have to select λ to make the $Q(x)$ sum to one, that is

$$e^{-(1+\lambda)} = \frac{1}{\sum_x P(x)e^{Q^*(x)}}, \tag{38}$$

meaning that the optimizing $Q(x)$ is given by

$$Q(x) = \frac{P(x)e^{Q^*(x)}}{\sum_{\tilde{x}} P(\tilde{x})e^{Q^*(\tilde{x})}}. \tag{39}$$

Plugging this particular choice of $Q(x)$ back into our main expression, we find

$$f^*(Q^*) = \max_Q \left\{ \left(\sum_{x\in\mathcal{X}} Q^*(x)Q(x) \right) - f(Q) \right\} \tag{40}$$

$$= \sum_{x\in\mathcal{X}} Q^*(x) \frac{P(x)e^{Q^*(x)}}{\sum_{\tilde{x}} P(\tilde{x})e^{Q^*(\tilde{x})}} - \sum_{x\in\mathcal{X}} \frac{P(x)e^{Q^*(x)}}{\sum_{\tilde{x}} P(\tilde{x})e^{Q^*(\tilde{x})}} \log \left(\frac{e^{Q^*(x)}}{\sum_{\tilde{x}} P(\tilde{x})e^{Q^*(\tilde{x})}} \right) \tag{41}$$

$$= \sum_{x\in\mathcal{X}} Q^*(x) \frac{P(x)e^{Q^*(x)}}{\sum_{\tilde{x}} P(\tilde{x})e^{Q^*(\tilde{x})}} - \sum_{x\in\mathcal{X}} \frac{P(x)e^{Q^*(x)}}{\sum_{\tilde{x}} P(\tilde{x})e^{Q^*(\tilde{x})}} Q^*(x)$$

$$+ \sum_{x\in\mathcal{X}} \frac{P(x)e^{Q^*(x)}}{\sum_{\tilde{x}} P(\tilde{x})e^{Q^*(\tilde{x})}} \log \left(\sum_{\tilde{x}} P(\tilde{x})e^{Q^*(\tilde{x})} \right) \tag{42}$$

$$= \log \left(\sum_{\tilde{x}} P(\tilde{x})e^{Q^*(\tilde{x})} \right). \tag{43}$$

This includes the logarithm of the moment-generating function of P as a special case (select $Q^*(x) = \lambda x$).

(b) Let $\mathcal{X} = \{x : P(x) > 0\}$. Then,

$$D(Q\|P) = \int_{x\in\mathcal{X}} Q(x) \log \frac{Q(x)}{P(x)} dx \tag{44}$$

$$= -h(Q) - \int_{x\in\mathcal{X}} Q(x) \log P(x) dx \tag{45}$$

$$= -h(Q) - \int_{x\in\mathcal{X}} Q(x) \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} \right) dx \tag{46}$$

$$= -h(Q) - \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) + \int_{x\in\mathcal{X}} Q(x) \frac{x^2}{2\sigma^2} dx \tag{47}$$

$$= -h(Q) + \frac{1}{2} \log \left(2\pi\sigma^2\right) + \frac{\mathbb{E}_Q[X^2]}{2\sigma^2} \tag{48}$$

$$= -h(Q) + \frac{1}{2} \log \left(2\pi\sigma^2\right) + \frac{1}{2} \tag{49}$$

$$= -h(Q) + \frac{1}{2} \log \left(2\pi\sigma^2\right) + \frac{1}{2} \log e \tag{50}$$

$$= -h(Q) + \frac{1}{2} \log \left(2\pi e\sigma^2\right), \tag{51}$$

where we recognize the second summand to be exactly the differential entropy of the Gaussian distribution with variance σ^2.

Alternatively, since we assume that second moments are equal, we could have observed that

$$D(Q\|P) = -h(Q) - \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) + \int_{x\in\mathcal{X}} Q(x) \frac{x^2}{2\sigma^2} dx \tag{52}$$

$$= -h(Q) - \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) + \int_{x\in\mathcal{X}} P(x) \frac{x^2}{2\sigma^2} dx \tag{53}$$

$$= -h(Q) - \int_{x\in\mathcal{X}} P(x) \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} \right) dx \tag{54}$$

$$= -h(Q) - \int_{x\in\mathcal{X}} P(x) \log P(x) dx, \tag{55}$$

where the second summand is precisely the entropy of $P(x)$.

Problem 6: Moments and Rényi

Suppose G is an integer valued random variable taking values in the set $\{1,\dots,K\}$. 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 $\{1,\dots,K\}$, 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 $U_1,\dots,U_n$ are i.i.d., each with distribution p. Suppose we try to determine the value of $X = (U_1,\dots,U_n)$ 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.]

Solution 6.

1. Simplifying the right hand side of the equation, we can get
$$\sum_i q_i \exp\Big[\log \frac{p_i i^\rho}{q_i}\Big] = \sum_i q_i \frac{p_i i^\rho}{q_i} = \sum_i p_i i^\rho = \mathbb{E}[G^\rho]$$

2. By Jensen's inequality and $\exp(x)$ is a convex function
$$\begin{aligned}\mathbb{E}[G^\rho] = \sum_i q_i \exp\Big[\log \frac{p_i i^\rho}{q_i}\Big] &\geq \exp\Big[\sum_i q_i \log \frac{p_i i^\rho}{q_i}\Big] \\ &= \exp\Big[\sum_i q_i \log \frac{p_i}{q_i} + \rho \sum_i q_i \log i\Big] \\ &= \exp\Big[-D(q\|p) + \rho \sum_i q_i \log i\Big]\end{aligned}$$

3.

$$\begin{aligned}\sum_i q_i \log i &= \sum_i q_i(\log \frac{1}{q_i} - \log \frac{1}{iq_i}) \\ &= H(q) - \sum_i q_i \log \frac{1}{iq_i} \\ &\geq H(q) - \log \sum_i \frac{1}{i}\end{aligned}$$

where the last inequality is obtained by apply Jensen's inequality on concave function $\log(x)$.

4. Using previous results, we have

$$\begin{aligned}\mathbb{E}[G^\rho] &\geq \exp\Big[-D(q\|p) + \rho\sum_i q_i \log i\Big] \\ &\geq \exp\Big[-D(q\|p) + \rho(H(q) - \log\sum_i \frac{1}{i})\Big] \\ &= \exp\Big[\rho H(q) - D(q\|p) - \rho\log\sum_i \frac{1}{i}\Big] \\ &= \exp\Big[\rho H(q) - D(q\|p) - \rho\log(1+\ln K)\Big] \\ &= (1+\ln K)^{-\rho}\exp\Big[\rho H(q) - D(q\|p)\Big]\end{aligned}$$

5. Since U_i's are i.i.d with distribution p, $X = (U_1, \ldots, U_n)$ is the joint distribution p^n. If each U_i has K distinct values, then X has K^n values. Thus, $G_n \in \{1, \ldots, K^n\}$

   Recall from Homework 2 Problem 6 that

$$\max_q \rho H(q) - D(q\|p_X) = \rho H_{1/(1+\rho)}(p_X) = \rho\sum_{i=1}^n H_{1/(1+\rho)}(p_{U_i}) = n\rho H_{1/(1+\rho)}(p)$$

   Since the result of Part 4 holds for any q, it also holds for the q which maximizes $\rho H(q) - D(q\|p_X)$. Hence, we have

$$\begin{aligned}\liminf_n \frac{1}{n\rho}\log\mathbb{E}[G_n^\rho] &\geq \liminf_n \max_q \frac{1}{n\rho}\log\Big((1+\ln K^n)^{-\rho}\exp\Big[\rho H(q) - D(q\|p_X)\Big]\Big) \\ &= \liminf_n \frac{1}{n\rho}[-\rho\log(1+\ln K^n) + \max_q \rho H(q) - D(q\|p_X)] \\ &= \liminf_n -\frac{1}{n}\log(1+n\ln K) + H_{1/(1+\rho)}(p) \\ &= H_{1/(1+\rho)}(p)\end{aligned}$$

   In the last step, $\liminf_n -\frac{1}{n}\log(1+n\ln K) = 0$.

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 $A = a_1 + a_2$, $B = b_1 + b_2$,

$$b_1 f(a_1/b_1) + b_2 f(a_2/b_2) \geq B f(A/B);$$

and that in general, for any non-negative $a_1, \dots, a_k$, $b_1, \dots, b_k$, and $A = \sum_i a_i$, $B = \sum_i b_i$, we have

$$\sum_i b_i f(a_i/b_i) \geq B f(A/B).$$

[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. $\sum_u (p(u) - q(u))^2/q(u)$

Solution 7. *(a)* Since f is convex, for any $\lambda \in [0,1]$ and any $x_1, x_2 >$ we have

$$\lambda f(x_1) + (1-\lambda) f(x_2) \geq f(\lambda x_1 + (1-\lambda) x_2) \tag{56}$$

By substitution $x_1 = a_1/b_1$, $x_2 = a_2/b_2$ and $\lambda = b_1/(b_1+b_2)$:

$$\frac{b_1}{b_1+b_2} f(\frac{a_1}{b_1}) + (1 - \frac{b_1}{b_1+b_2}) f(\frac{a_2}{b_2}) \geq f(\frac{b_1}{b_1+b_2}\frac{a_1}{b_1} + (1 - \frac{b_1}{b_1+b_2})\frac{a_2}{b_2}) \tag{57}$$

$$\Leftrightarrow b_1 f(\frac{a_1}{b_1}) + b_2 f(\frac{a_2}{b_2}) \geq B f(A/B) \tag{58}$$

Let $A_k = \sum_{i=1}^k a_i$, $B_k = \sum_{i=1}^k b_i$. As we have proved that the following inequality holds for $k = 1, 2$:

$$\sum_{i=1}^k b_i f(a_i/b_i) \geq B_k f(A_k/B_k). \tag{59}$$

We assume that it also holds for $k = n$. For $k = n+1$, we have

$$\begin{aligned} \sum_{i=1}^{n+1} b_i f(a_i/b_i) &= \sum_{i=1}^{n} b_i f(a_i/b_i) + b_{n+1} f(a_{n+1}/b_{n+1}) && (60)\\ &\geq B_n f(A_n/B_n) + b_{n+1} f(a_{n+1}/b_{n+1}) && (61)\\ &\geq B_{n+1} f(A_{n+1}/B_{n+1}) && (62) \end{aligned}$$

By induction, for all any non-negative k, we have

$$\sum_{i=1}^{k} b_i f(a_i/b_i) \geq B_k f(A_k/B_k). \tag{63}$$

(b) $D_f(p\|q) = \sum_u q(u) f(p(u)/q(u)) \geq (\sum_u q(u)) f(\frac{\sum_u p(u)}{\sum_u q(u)}) = 1 f(1) = 0$.

(c)

$$\begin{aligned} D_f(p\|q) = \sum_u q(u) f(p(u)/q(u)) &= \sum_u \sum_v W(v|u) q(u) f(p(u)/q(u)) && (64)\\ &= \sum_u \sum_v W(v|u) q(u) f(W(v|u)p(u)/(W(v|u)q(u))) && (65)\\ &\geq \sum_v (\sum_u W(v|u) q(u)) f\left(\frac{\sum_u W(v|u)p(u)}{\sum_u W(v|u)q(u)}\right) && (66)\\ &= \sum_v \tilde{q}(v) f(\tilde{p}(v)/\tilde{q}(v)) && (67)\\ &= D_f(\tilde{p}\|\tilde{q}) && (68) \end{aligned}$$

(d)

i. $D(p\|q) := \sum_u p(u) \log(p(u)/q(u)) = \sum_u q(u) \frac{p(u)}{q(u)} \log \frac{p(u)}{q(u)}$. So $f(t) = t \log t$.

ii. $R(p\|q) := D(q\|p) = \sum_u p(u) \log(p(u)/q(u)) = \sum_u p(u)(-\log(q(u)/p(u)))$. So $f(t) = -\log t$.

iii. $1 - \sum_u \sqrt{p(u)q(u)} = \sum_u q(u) \left(1 - \sqrt{p(u)/q(u)}\right)$. So $f(t) = 1 - \sqrt{t}$.

iv. $\|p - q\|_1 = \sum_u |p(u) - q(u)| = \sum_u q(u)|p(u)/q(u) - 1|$. So $f(t) = |t - 1|$.

v. $\sum_u (p(u) - q(u))^2/q(u) = \sum_u q(u)(p(u)/q(u) - 1)^2$. So $f(t) = (t - 1)^2$.