Tsinghua-Berkeley Shenzhen Institute
Information Theory and Statistical Learning
Fall 2020

Problem Set 2

Issued: Monday 28th September, 2020 **Due:** Monday 12th October, 2020

Notations: We use Bern(p) to denote the Bernoulli distribution with the parameter p, and use Binom(n, p) to denote the binomial distribution with parameters n and p.

2.1. Please use Chain Rule for mutual information to derive $I(X_1, \ldots, X_n; Y_1, \ldots, Y_m)$.

Solution: The answer is based on the fact that $I(X;Y) = I(Y;X)$.

$$\begin{aligned} I(X_1, \ldots, X_n; Y_1, \ldots, Y_m) &= \sum_{i=1}^{n} I(X_i; Y_1, \ldots, Y_m | X_1, \ldots, X_{i-1}) \\ &= \sum_{i=1}^{n} \left(\sum_{j=1}^{m} I(X_i; Y_j | X_1, \ldots, X_{i-1}, Y_1, \ldots, Y_{j-1}) \right) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{m} I(X_i; Y_j | X_1, \ldots, X_{i-1}, Y_1, \ldots, Y_{j-1}) \end{aligned}$$

2.2. *Conditional mutual information vs. unconditional mutual information.* Give examples of joint random variables X, Y, and Z such that

(a) $I(X;Y|Z) < I(X;Y)$.

(b) $I(X;Y|Z) > I(X;Y)$

Solution: It is *Problem 2.6* in Cover's book.

(a) The last corollary to Theorem 2.8.1 in the text states that if $X \to Y \to Z$ that is, if $p(x, y|z) = p(x|z)p(y|z)$ then, $I(X;Y) \leq I(X;Y|Z)$. Equality holds if and only if $I(X;Z) = 0$ or X and Z are independent.
A simple example of random variables satisfying the inequality conditions above is, X is a fair binary random variable and $Y = X$ and $Z = Y$. In this case,

$$I(X;Y) = H(X) - H(X|Y) = H(X) = 1$$

and,

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) = 0.$$

So that $I(X;Y|Z) < I(X;Y)$

(b) This example is also given in the text. Let X, Y be independent fair binary random variables and let $Z = X + Y$. In this case we have that,

$$I(X;Y) = 0$$

and,

$$I(X;Y|Z) = H(X|Z) = 1/2.$$

So $I(X;Y|Z) > I(X;Y)$. Note that in this case X, Y, Z are not markov.

2.3. *Information measures.* Suppose $Z_1, \dots, Z_n$ are i.i.d. Bern$\left(\frac{1}{2}\right)$ random variables, and let $X_A \triangleq (Z_i)_{i \in A}$ be the random vector consisting of the bits with indices in A. Prove that

(a) For all non-empty $A \subset \{1, \dots, n\}$, we have $H(X_A) = |A|$.

(b) For all non-empty $A_1, A_2 \subset \{1, \dots, n\}$, we have

$$H(X_{A_1}, X_{A_2}) = |A_1 \cup A_2|, \tag{1a}$$
$$H(X_{A_1} | X_{A_2}) = |A_1 \setminus A_2|, \tag{1b}$$
$$I(X_{A_1}; X_{A_2}) = |A_1 \cap A_2|. \tag{1c}$$

Solution:

(a) Since all Z_i's are independent, we have

$$H(X_A) = \sum_{i \in A} H(Z_i) = \sum_{i \in A} 1 = |A|.$$

(b) To obtain (1a) note that

$$H(X_{A_1}, X_{A_2}) = \sum_{i \in A_1 \cup A_2} H(Z_i) = |A_1 \cup A_2|.$$

The other two equalities can be obtained using

$$H(X_{A_1} | X_{A_2}) = H(X_{A_1}, X_{A_2}) - H(X_{A_2})$$

and

$$I(X_{A_1}; X_{A_2}) = H(X_{A_1}) + H(X_{A_2}) - H(X_{A_1}, X_{A_2}),$$

respectively.

2.4. Let (X, Y) be uniformly distributed in the unit l_p-ball $B_p \triangleq \{(x, y) : |x|^p + |y|^p \leq 1\}$, where $p \in (0, \infty)$. Also define the l_∞-ball $B_\infty \triangleq \{(x, y) : |x| \leq 1, |y| \leq 1\}$.

(a) Are X and Y independent for $p = 1$?

(b) Compute $I(X; Y)$ for $p = \frac{1}{2}$, $p = 1$ and $p = \infty$.

(c) What do you think $I(X; Y)$ converges to as $p \to 0$. Explain it.

Solution:

(a) Not independent. Obviously, $p_{X|Y}(x|0) \neq f_{X|Y}(x|1)$.

(b) Due to the symmetry of X and Y, $I(X; Y) = 2H(X) - H(X, Y)$. $H(X, Y) = \log S_p$, where S_p is the area of the unit l_p-ball.

$$S_p = 4 \int_0^1 (1 - x^p)^{\frac{1}{p}} \, \mathrm{d}x$$

Let $q = \frac{1}{p}$ and $x = t^q$. Then,

$$S_p = 4q\int_0^1 (1-t)^q t^{q-1}\mathrm{d}t = 4q\frac{\Gamma(q+1)\Gamma(q)}{\Gamma(2q+1)} = 4\frac{\Gamma(q+1)\Gamma(q+1)}{\Gamma(2q+1)}.$$

The marginal distribution of X is $p_X(x) = \frac{2}{S_p}\left(1-|x|^p\right)^{\frac{1}{p}}$. Then,

$$\begin{aligned}
H(X) &= -\int_0^1 \frac{4}{S_p}\left(1-x^p\right)^{\frac{1}{p}}\log\left(\frac{2}{S_p}\left(1-x^p\right)^{\frac{1}{p}}\right)\mathrm{d}x \\
&= \log\frac{S_p}{2} - \int_0^1 \frac{4q}{S_p}\left(1-x^{1/q}\right)^q \log\left(1-x^{1/q}\right)\mathrm{d}x.
\end{aligned}$$

Let $1 - x^{1/q} = s$. Then, $x = (1-s)^q$ and

$$H(X) = \log\frac{S_p}{2} + \int_0^1 \frac{4q}{S_p} s^q \log s \mathrm{d}(1-s)^q.$$

We need the Digamma function here. Please see Beta Function and Digamma Function. When $q \le 1$,

$$\begin{aligned}
H(X) &= \log\frac{S_p}{2} - \int_0^1 \frac{4q^2}{S_p} s^q(1-s)^{q-1}\log s\mathrm{d}s \\
&= \log\frac{S_p}{2} - \frac{4q^2}{S_p}\cdot \mathrm{B}(q+1,q)(\psi(q+1)-\psi(2q+1)) \\
&= \log\frac{S_p}{2} + q\sum_{i=q+1}^{2q}\frac{1}{i}
\end{aligned}$$

So let's see the following p's.

i) $p = 1/2 \to q = 2 \to S_{1/2} = \frac{2}{3}$

$$I(X;Y) = 2\log\frac{1}{3} + 4\times\left(\frac{1}{3}+\frac{1}{4}\right) - \log\frac{2}{3} = \frac{7}{3} - \log 2 - \log 3$$

ii) $p = 1 \to q = 1 \to S_1 = 2$

$$\begin{aligned}
I(X;Y) &= 2\log 1 + 2\times\frac{1}{2} - \log 2 \\
&= 1 - \log 2
\end{aligned}$$

iii) $p = \infty \to q = 0 \to X$ and Y are indepedent, So $I(X;Y) = 0$

(c) $I(X;Y)$ converges to $+\infty$. Here is the insights. When $p \to 0$, it means if we know Y, we can more precisely predict the opposite X, because the boundary will getting closer to the x and y axis. It means the common information will account for a larger part of the entropy. Then we claim that $I(X;Y)$ will also goes to infinity. A detailed explanation is the following.

$p \to 0$ means $q \to +\infty$, and $S_p = 4\frac{\Gamma(q+1)\Gamma(q+1)}{\Gamma(2q+1)} = \frac{4q!q!}{(2q)!} \to 0$

$$I(X;Y) = \log \frac{q!q!}{(2q)!} + 2q \sum_{i=q+1}^{2q} \frac{1}{i}$$

We need 2 approximations here. The first is $\lim_{q\to\infty} \sum_{i=q+1}^{2q} \frac{1}{i} = \ln 2$. It is a common conclusion derived by integrals. The second is Stirling's approximation that $n! \sim \sqrt{2\pi n}(\frac{n}{e})^n$. Then,

$$\log \frac{q!q!}{(2q)!} \sim \log \left(\frac{\sqrt{\pi q}}{2^{2q}} \right) = -2q \ln 2 + \frac{1}{2} \log q + \frac{1}{2} \log \pi.$$

Therefore,

$$I(X;Y) \sim \frac{1}{2} \log q + \frac{1}{2} \log \pi \to +\infty$$

2.5. Let $\mathcal{N}(\boldsymbol{m}, \boldsymbol{\Sigma})$ be the Gaussian distribution on $\mathbb{R}^n$ with mean $\boldsymbol{m} \in \mathbb{R}^n$ and covariance matrix $\boldsymbol{\Sigma}$.

(a) Under what conditions on $\boldsymbol{m}_0, \boldsymbol{\Sigma}_0, \boldsymbol{m}_1, \boldsymbol{\Sigma}_1$ is

$$D\left(\mathcal{N}(\boldsymbol{m}_1, \boldsymbol{\Sigma}_1) \| \mathcal{N}(\boldsymbol{m}_0, \boldsymbol{\Sigma}_0)\right) < \infty \tag{2}$$

(b) Compute $D\left(\mathcal{N}(\boldsymbol{m}, \boldsymbol{\Sigma}) \| \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)\right)$, where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix.

(c) Compute $D\left(\mathcal{N}(\boldsymbol{m}_1, \boldsymbol{\Sigma}_1) \| \mathcal{N}(\boldsymbol{m}_0, \boldsymbol{\Sigma}_0)\right)$ for a non-singular $\boldsymbol{\Sigma}_0$.

Solution: For convenience, we take the natural logarithm in the definition of K-L divergence.

(a) The condition is called *absolute continuity of measures*. In this case, that is to avoid $\log \infty$ defined on a non-zero measure. So $\boldsymbol{\Sigma}_0$ should be non-singular.

(b) The last 2 answer is shown together. You can find the result in each textbook.

$$\begin{aligned}
&D\left(\mathcal{N}(\boldsymbol{m}_1, \boldsymbol{\Sigma}_1) \| \mathcal{N}(\boldsymbol{m}_0, \boldsymbol{\Sigma}_0)\right) \\
&= \mathbb{E}_{\underline{X}\sim\mathcal{N}(\boldsymbol{m}_1,\boldsymbol{\Sigma}_1)} \left[\log \frac{\sqrt{|2\pi\boldsymbol{\Sigma}_0|} \exp(-\frac{1}{2}(\underline{X} - \boldsymbol{m}_1)^T \boldsymbol{\Sigma}_1^{-1} (\underline{X} - \boldsymbol{m}_1))}{\sqrt{|2\pi\boldsymbol{\Sigma}_1|} \exp(-\frac{1}{2}(\underline{X} - \boldsymbol{m}_0)^T \boldsymbol{\Sigma}_0^{-1} (\underline{X} - \boldsymbol{m}_0))} \right] \\
&= \frac{1}{2} \log \frac{|\boldsymbol{\Sigma}_0|}{|\boldsymbol{\Sigma}_1|} - \frac{1}{2} \mathbb{E}_{\underline{X}\sim\mathcal{N}(\boldsymbol{m}_1,\boldsymbol{\Sigma}_1)} \left[(\underline{X} - \boldsymbol{m}_1)^T \boldsymbol{\Sigma}_1^{-1} (\underline{X} - \boldsymbol{m}_1) \right] \\
&\quad + \frac{1}{2} \mathbb{E}_{\underline{X}\sim\mathcal{N}(\boldsymbol{m}_1,\boldsymbol{\Sigma}_1)} \left[(\underline{X} - \boldsymbol{m}_0)^T \boldsymbol{\Sigma}_0^{-1} (\underline{X} - \boldsymbol{m}_0) \right] \\
&= \frac{1}{2} \log \frac{|\boldsymbol{\Sigma}_0|}{|\boldsymbol{\Sigma}_1|} - \frac{1}{2} n + \frac{1}{2} \operatorname{tr}(\boldsymbol{\Sigma}_1 \boldsymbol{\Sigma}_0^{-1}) + \frac{1}{2} (\boldsymbol{m}_1 - \boldsymbol{m}_0)^T \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{m}_1 - \boldsymbol{m}_0)
\end{aligned}$$

2.6. There are two probability distribution P and Q over a finite alphabet $\mathcal{X}$ with cardinality k. Let us use $P_1 \geq P_2 \geq \cdots \geq P_k$ and $Q_1 \geq Q_2 \geq \cdots \geq Q_k$ to denote the non-increasing ordering of p.m.f P and Q respectively ($\sum_{i=1}^k P_i = \sum_{i=1}^k Q_i = 1$). We say that P is *more uniform* then Q if

$$\forall l \in [1:k], \sum_{i=1}^{l} P_i \leq \sum_{i=1}^{l} Q_i \tag{3}$$

In this problem, we would like to prove that if P is more uniform then Q in the sense of (3), then

$$H(P) \geq H(Q) \tag{4}$$

(a) Prove that for convex function $f(\cdot)$, $\sum_{i=1}^k f(P_i) \leq \sum_{i=1}^k f(Q_i)$.

(b) Use (a) to prove (4)

Solution:
a) First we prove a property of convex function:

Lemma 1. Let $f : \mathbb{R} \to \mathbb{R}$ be a convex function. For a fixed x_1, the slope

$$R(x_1, x_2) = \frac{f(x_1) - f(x_2)}{x_1 - x_2}$$

is monotonically non-decreasing for x_2 as $x_2 < x_1$. Similarly, for a fixed x_2, $R(x_1, x_2)$ is monotonically non-decreasing for x_1 as $x_1 > x_2$.

Proof. For the first case (fixed x_1), consider any $x_1, x_2, x_2^{'}$, such that $x_2 < x_2^{'} < x_1$, we can write $x_2^{'}$ as a convex combination of x_1 and x_2:

$$x_2^{'} = \theta x_2 + (1-\theta)x_1,\ 0 < \theta = \frac{x_1 - x_2^{'}}{x_1 - x_2} < 1$$

By definition of convex function, we have

$$f(x_2^{'}) \leq \theta f(x_2) + (1-\theta) f(x_1)$$

that is,

$$(x_1 - x_2) f(x_2^{'}) \leq (x_1 - x_2^{'}) f(x_2) + (x_2^{'} - x_2) f(x_1)$$
$$(x_1 - x_2^{'}) f(x_1) + (x_1 - x_2) f(x_2^{'}) \leq (x_1 - x_2^{'}) f(x_2) + (x_1 - x_2) f(x_1)$$
$$\frac{f(x_1) - f(x_2)}{x_1 - x_2} \leq \frac{f(x_1) - f(x_2^{'})}{x_1 - x_2^{'}}$$

The proof of fixed x_2 case is similar. Or you can just use the fact that $f(x)$ is also convex. $\square$

Now come back to the original problem. Well prove a slightly stronger version: instead of limiting P, Q be probability distributions, now we only require them to have the same sum,

$$\sum_{i=1}^{k} P_i = \sum_{i=1}^{k} Q_i \tag{6}$$

If $P_i = Q_i$ for all $1 \leq i \leq k$, then the problem holds with equality. So we can assume there exists at least one i that $P_i \neq Q_i$.
If there exists an i that $P_i = Q_i$, we can remove that term from both sequences, and the property (1) and (6) still holds. So we can assume that $P_i \neq Q_i$ for all i.
For $1 \leq i \leq k$, define

$$A_0 = B_0 = 0$$

$$A_i = \sum_{j=1}^{i} Q_j,\ B_i = \sum_{j=1}^{i} P_j$$

$$Q_i^{'} = \max(Q_i, P_i), P_i^{'} = \min(Q_i, P_i)$$

$$c_i = \frac{f(Q_i) - f(P_i)}{Q_i - P_i}$$

$$c_i^{'} = \frac{f(Q_i^{'}) - f(P_i^{'})}{Q_i^{'} - P_i^{'}}$$

By (1), we have $A_i \geq B_i$ for $0 \leq i \leq k$. Swapping Q_i and P_i doesnt change the slope, hence $c_i = c_i^{'}$. For $1 \leq i \leq k-1$,

$$P_i^{'} < Q_i^{'},\ P_{i+1}^{'} < Q_{i+1}^{'}$$

$$Q_i^{'} = \max(Q_i, P_i) \geq \max(Q_{i+1}, P_{i+1}) = Q_{i+1}^{'}$$

$$P_i^{'} = \min(Q_i, P_i) \geq \min(Q_{i+1}, P_{i+1}) = P_{i+1}^{'}$$

Now by Lemma 1,

$$c_i = c_i^{'} = \frac{f(Q_i^{'}) - f(P_i^{'})}{Q_i^{'} - P_i^{'}} \geq \frac{f(Q_i^{'}) - f(P_{i+1}^{'})}{Q_i^{'} - P_{i+1}^{'}} \geq \frac{f(Q_{i+1}^{'}) - f(P_{i+1}^{'})}{Q_{i+1}^{'} - P_{i+1}^{'}} = c_{i+1}^{'} = c_{i+1}$$

Using these properties, we have

$$\begin{aligned}
\sum_{i=1}^{k}(f(Q_i) - f(P_i)) &= \sum_{i=1}^{k} c_i(Q_i - P_i) \\
&= \sum_{i=1}^{k} c_i(A_i - B_i) - \sum_{i=1}^{k} c_i(A_{i-1} - B_{i1}) \\
&= \sum_{i=1}^{k} c_i(A_i - B_i) - \sum_{i=0}^{k-1} c_{i+1}(A_i - Bi) \\
&= c_k(A_k - B_k) - c_1(A_0 - B_0) + \sum_{i=1}^{k-1}(c_i - c_{i+1})(A_i - B_i) \\
&= \sum_{i=1}^{k-1}(c_i - c_{i+1})(A_i - B_i) \geq 0
\end{aligned}$$

The second last line is by (6), and the last line is because $c_i \geq c_{i+1}$ and $A_i \geq B_i$. Hence we get

$$\sum_{i=1}^{k}(f(Q_i) - f(P_i)) \geq 0$$

that is

$$\sum_{i=1}^{k} f(P_i) \leq \sum_{i=1}^{k} f(Q_i)$$

b) The entropies are $H(P) = \sum P_i \log(P_i), H(Q) = \sum Q_i \log(Q_i)$. Let $f(x) = x\log(x)$, then $f(x)$ is convex on $(0, \infty)$. By a), $\sum_{i=1}^{k} P_i \log(P_i) \leq \sum_{i=1}^{k} Q_i \log(Q_i)$. So, $H(P) \geq H(Q)$.
Comments: Despite of such a long proof, the property is described as Schur-convexity, named after Issai Schur. See https://en.wikipedia.org/wiki/Schur-convex_function

2.7. *Total correlation.* For a given set of n random variables $X_1, \dots, X_n$, the total correlation $C(X_1, \dots, X_n)$ is defined as the K-L divergence from the joint distribution to the product distribution, i.e.,

$$C(X_1, \dots, X_n) \triangleq D\left(P_{X^n} \middle\| \prod_{i=1}^{n} P_{X_i}\right).$$

(a) Prove that

$$\begin{aligned} C(X_1, \dots, X_n) &= \sum_{i=1}^{n} H(X_i) - H(X^n) && (5a) \\ &= \sum_{i=1}^{n-1} I(X^i; X_{i+1}). && (5b) \end{aligned}$$

(b) When will the total correlation be zero?

Solution:

(a) ■

(b) Since the K-L divergence is zero iff the two distributions are identical, we know that $X_1, \dots, X_n$ are independent.

2.8. *Divergence of order statistics.* Given $x^n = (x_1, \dots, x_n) \in \mathbb{R}^n$, let $x_{(1)} \leq \cdots \leq x_{(n)}$ denote the ordered entries. Let P, Q be distributions on $\mathbb{R}$ and $P_{X^n} = P^n, Q_{X^n} = Q^n$.

(a) Prove that

$$D(P_{X_{(1)} \cdots X_{(n)}} \| Q_{X_{(1)} \cdots X_{(n)}}) = nD(P\|Q). \quad (6)$$

(b) Show that

$$D(\mathrm{Binom}(n,p)\|\,\mathrm{Binom}(n,q)) = nD(\mathrm{Bern}(p)\|\,\mathrm{Bern}(q)). \tag{7}$$

Solution:

(a) Compute the joint distribution $P_{X_{(1)}\dots X_{(n)}}$ and $Q_{X_{(1)}\dots X_{(n)}}$, and then use the definition. Note that

- $P_{X_{(1)}\dots X_{(n)}} \neq P_{X^n}$, although we can verify that

$$D(P_{X^n}\|Q_{X^n}) = D(P_{X_{(1)}\dots X_{(n)}}\|Q_{X_{(1)}\dots X_{(n)}}) = nD(P\|Q).$$

- The case for discrete random variables is different from the case where X_i's are continuous random variable.

(b) You can directly apply the conclusion of (a). Suppose that $X_i \overset{\text{i.i.d.}}{\sim} \mathrm{Bern}(p)$, then there exists a one-to-one mapping between $(X_{(1)},\dots,X_{(n)})$ and $Y \triangleq \sum_{i=1}^n X_i \sim \mathrm{Binom}(n,p)$. Therefore, we have

$$\begin{aligned} D(\mathrm{Binom}(n,p)\|\,\mathrm{Binom}(n,q)) &= D(P_Y\|Q_Y) = D(P_{X_{(1)}\dots X_{(n)}}\|Q_{X_{(1)}\dots X_{(n)}}) \\ &= nD(\mathrm{Bern}(p)\|\,\mathrm{Bern}(q)), \end{aligned}$$

where we have assumed that $P_X = \mathrm{Bern}(p), Q_X = \mathrm{Bern}(q)$.