Tsinghua-Berkeley Shenzhen Institute
INFORMATION THEORY AND STATISTICAL LEARNING
Fall 2020

Homework 2

Hanmo Chen 2020214276 September 29, 2020

- **Acknowledgments:**

  For the Gamma function and its series extensions, I refer to the website(`https://www.wolframalpha.com/`). And for the proof of Karamata's Inequality, I refer to the wikipedia(`https://en.wikipedia.org/wiki/Karamata's_inequality`). Some matrix tricks used in deriviation comes from the Matrix Cookbook(`http://matrixcookbook.com/`)

- **Collaborators:** None

- *I certify that all solutions are entirely in my words and that I have not looked at another student's solutions. I have credited all external sources in this write up.*

  Hanmo Chen

2.1. First, we have Chain Rule for mutual information

$$
\begin{aligned}
I\left(X_{1}, X_{2}, \cdots, X_{n} ; Y\right) &= \sum_{i=1}^{n} I\left(X_{i} ; Y \mid X_{i-1}, X_{i-2}, \cdots, X_{1}\right) \\
I\left(X ; Y_{1}, Y_{2}, \cdots, Y_{n}\right) &= \sum_{i=1}^{n} I\left(X ; Y_{i} \mid Y_{i-1}, Y_{i-2}, \cdots, Y_{1}\right)
\end{aligned} \tag{1}
$$

Using this Chain Rule twice,

$$
\begin{aligned}
I\left(X_{1}, \ldots, X_{n} ; Y_{1}, \ldots, Y_{m}\right) &= \sum_{j=1}^{m} I\left(X_{1}, X_{2}, \cdots, X_{n} ; Y_{j} \mid Y_{j-1}, Y_{j-2}, \cdots, Y_{1}\right) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{m} I\left(X_{i} ; Y_{j} \mid X_{i-1}, X_{i-2}, \cdots, X_{1}, Y_{j-1}, Y_{j-2}, \cdots, Y_{1}\right)
\end{aligned} \tag{2}
$$

2.2. Examples:

(a) $I(X;Y \mid Z) < I(X;Y)$
Let $Z = X$ then $I(X;Y \mid Z) = H(X \mid Z) - H(X,Y \mid Z) = 0$. But for X, Y that are not independent, $I(X;Y) > 0 = I(X;Y \mid Z)$

(b) $I(X;Y \mid Z) > I(X;Y)$
Use the example illustrated in class. let X and Y be independent Bernoulli random variables with $p = 0.5$, and let $Z = X + Y$. Then $I(X;Y) = 0$, but
$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) = H(X|Z) = 0.5$ bit.

2.3. *Information Measures.*

(a) Denote $n = |A|$, because $Z_1, Z_2, \cdots, Z_n$ are i.i.d Bern($\frac{1}{2}$) random variables, so $X_A = (Z_i)_{i \in A}$ is a n-dimension random vector having 2^n possible points with equal probability,

$$H(X_A) = 2^n * (-\frac{1}{2^n} \log \frac{1}{2^n}) = n = |A| \text{ (bits)} \tag{3}$$

(b) Denote $A = A_1 \cup A_2, A^* = A_1 \cap A_2, A_1^* = A_1 \backslash A_2, A_2^* = A_2 \backslash A_1$, so $X_{A_1} = (X_{A^*}, X_{A_1^*}), X_{A_2} = (X_{A^*}, X_{A_2^*})$. Because $X_{A^*}, X_{A_1^*}, X_{A_2^*}$ are mutually independent,

$$\begin{aligned} H(X_{A_1}|X_{A_2}) &= H((X_{A^*}, X_{A_1^*})|(X_{A^*}, X_{A_2^*})) \\ &= H(X_{A_1^*}) = |A_1^*| = |A_1 \backslash A_2| \end{aligned} \tag{4}$$

Thus,

$$\begin{aligned} H(X_{A_1}, X_{A_2}) &= H(X_{A_2}) + H(X_{A_1}|X_{A_2}) \\ &= |A_2| + |A_1^*| = |A_1 \cup A_2| \end{aligned} \tag{5}$$

$$\begin{aligned} I(X_{A_1}; X_{A_2}) &= H(X_{A_2}) - H(X_{A_2}|X_{A_1}) \\ &= |A_2| - |A_2^*| = |A_1 \cap A_2| \end{aligned} \tag{6}$$

2.4. (a) No. $f_{X,Y}(x,y) = 0.5, |x| + |y| \leqslant 1$, and
$f_X(x) = 1 - |x|, x \in [-1,1], f_Y(y) = 1 - |y|, y \in [-1,1]$,
$f_{X,Y}(x,y) \neq f_X(x) f_Y(y)$ so X, Y are not independent.

(b) Because X and Y are symmetric, $H(X) = H(Y)$ and
$I(X;Y) = H(X) + H(Y) - H(X,Y) = 2H(X) - H(X,Y)$.
Denote the area of B_p as $S_p = 4 \int_0^1 (1 - |x|^p)^{1/p} dx$,
joint distribution of (X,Y) is $f_p(x,y) = \frac{1}{S_p}, |x|^p + |y|^p \leqslant 1$
the *pdf* of X is $f_p(x) = \frac{2}{S_p}(1 - |x|^p)^{1/p}, x \in [-1,1]$
So

$$I(X;Y) = 2H(X) - H(X,Y) = -2 \int_{-1}^{1} f_p(x) \log(f_p(x)) dx - \log S_p \tag{7}$$

- $p = \frac{1}{2}, S_p = \frac{2}{3}$, $I(X;Y) = \frac{7}{3} - \ln 3 - \ln 2 = 0.542$
- $p = 1, S_p = 2, I(X;Y) = 1 - \ln 2 = 0.307$
- $p = \infty$, X, Y are independent, $I(X;Y) = 0$

(c)

$$\lim_{p\to 0} I(X;Y) = \infty \tag{8}$$

Denote $n = 1/p$,

$$S_p = 4\int_0^1 (1 - x^{\frac{1}{n}})^n dx = 4n\int_0^1 t^n (1-t)^{n-1} dt = 4\frac{(\Gamma(n+1))^2}{\Gamma(2n+1)} \tag{9}$$

And

$$I_n(X;Y) = 2n(\sum_{i=n+1}^{2n} \frac{1}{i}) - \log\frac{(2n)!}{(n!)^2} \tag{10}$$

Because

$$\lim_{n\to\infty} \sum_{i=n+1}^{2n} \frac{1}{i} = \ln 2 \tag{11}$$

$$\log\frac{(2n)!}{(n!)^2} = \log\left(\frac{2^{2n}\Gamma\left(n+\frac{1}{2}\right)}{\Gamma(n+1)}\right) - \frac{\log(\pi)}{2} \tag{12}$$

And $\frac{\Gamma\left(n+\frac{1}{2}\right)}{\Gamma(n+1)} \sim \log(n)$ as $n \to \infty$, thus

$$\lim_{n\to\infty} I_n(X;Y) = \lim_{n\to\infty} \frac{1}{2}\log(n) = \infty \tag{13}$$

2.5. The density function of Gaussian distribution is

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{m})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{m})\right) \tag{14}$$

Thus,

$$\begin{aligned}
D\left(\mathcal{N}\left(\boldsymbol{m}_1, \boldsymbol{\Sigma}_1\right) \| \mathcal{N}\left(\boldsymbol{m}_0, \boldsymbol{\Sigma}_0\right)\right) &= \mathbb{E}_{\mathcal{N}(\boldsymbol{m}_1,\boldsymbol{\Sigma}_1)}[\log\left(\frac{p_1(\mathbf{x})}{p_0(\mathbf{x})}\right)] \\
&= \mathbb{E}_{\mathcal{N}(\boldsymbol{m}_1,\boldsymbol{\Sigma}_1)}\left[\frac{1}{2}\log\frac{|\boldsymbol{\Sigma}_0|}{|\boldsymbol{\Sigma}_1|} - \frac{1}{2}\left(\mathbf{x}-\boldsymbol{m}_1\right)^T\boldsymbol{\Sigma}_1^{-1}\left(\mathbf{x}-\boldsymbol{m}_1\right)\right. \\
&\quad \left. + \frac{1}{2}\left(\mathbf{x}-\boldsymbol{m}_0\right)^T\boldsymbol{\Sigma}_0^{-1}\left(\mathbf{x}-\boldsymbol{m}_0\right)\right] \\
&= \frac{1}{2}\log\frac{|\boldsymbol{\Sigma}_0|}{|\boldsymbol{\Sigma}_1|} - \frac{1}{2}\mathbb{E}_{\mathcal{N}(\boldsymbol{m}_1,\boldsymbol{\Sigma}_1)}\left[\left(\mathbf{x}-\boldsymbol{m}_1\right)^T\boldsymbol{\Sigma}_1^{-1}\left(\mathbf{x}-\boldsymbol{m}_1\right)\right] \\
&\quad + \frac{1}{2}\mathbb{E}_{\mathcal{N}(\boldsymbol{m}_1,\boldsymbol{\Sigma}_1)}\left[\left(\mathbf{x}-\boldsymbol{m}_0\right)^T\boldsymbol{\Sigma}_0^{-1}\left(\mathbf{x}-\boldsymbol{m}_0\right)\right]
\end{aligned} \tag{15}$$

For $\mathbb{E}_{\mathcal{N}(\boldsymbol{m}_1,\boldsymbol{\Sigma}_1)}\left[\left(\mathbf{x}-\boldsymbol{m}_1\right)^T\boldsymbol{\Sigma}_1^{-1}\left(\mathbf{x}-\boldsymbol{m}_1\right)\right]$,

$$
\begin{aligned}
\mathbb{E}_{\mathcal{N}(\boldsymbol{m}_1,\boldsymbol{\Sigma}_1)}\left[(\mathbf{x}-\boldsymbol{m}_1)^T\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{m}_1)\right] &= \operatorname{tr}\left(\mathbb{E}_{\mathcal{N}(\boldsymbol{m}_1,\boldsymbol{\Sigma}_1)}\left[(\mathbf{x}-\boldsymbol{m}_1)^T\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{m}_1)\right]\right) \\
&= \mathbb{E}_{\mathcal{N}(\boldsymbol{m}_1,\boldsymbol{\Sigma}_1)}\left[\operatorname{tr}\left((\mathbf{x}-\boldsymbol{m}_1)^T\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{m}_1)\right)\right] \\
&= \mathbb{E}_{\mathcal{N}(\boldsymbol{m}_1,\boldsymbol{\Sigma}_1)}\left[\operatorname{tr}\left((\mathbf{x}-\boldsymbol{m}_1)(\mathbf{x}-\boldsymbol{m}_1)^T\boldsymbol{\Sigma}_1^{-1}\right)\right] \\
&= \operatorname{tr}\left(\mathbb{E}_{\mathcal{N}(\boldsymbol{m}_1,\boldsymbol{\Sigma}_1)}\left[(\mathbf{x}-\boldsymbol{m}_1)(\mathbf{x}-\boldsymbol{m}_1)^T\right]\boldsymbol{\Sigma}_1^{-1}\right) \\
&= \operatorname{tr}\left(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_1^{-1}\right) = \operatorname{tr}(\boldsymbol{I}_n) = n
\end{aligned}
\tag{16}
$$

In the same way,

$$
\begin{aligned}
\mathbb{E}_{\mathcal{N}(\boldsymbol{m}_1,\boldsymbol{\Sigma}_1)}\left[(\mathbf{x}-\boldsymbol{m}_0)^T\boldsymbol{\Sigma}_0^{-1}(\mathbf{x}-\boldsymbol{m}_0)\right] &= \operatorname{tr}\left(\mathbb{E}_{\mathcal{N}(\boldsymbol{m}_1,\boldsymbol{\Sigma}_1)}\left[(\mathbf{x}-\boldsymbol{m}_0)(\mathbf{x}-\boldsymbol{m}_0)^T\right]\boldsymbol{\Sigma}_0^{-1}\right) \\
&= \operatorname{tr}\left(\left((\boldsymbol{m}_1-\boldsymbol{m}_0)(\boldsymbol{m}_1-\boldsymbol{m}_0)^T+\boldsymbol{\Sigma}_1\right)\boldsymbol{\Sigma}_0^{-1}\right) \\
&= \operatorname{tr}\left(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_0^{-1}\right)+(\boldsymbol{m}_1-\boldsymbol{m}_0)^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{m}_1-\boldsymbol{m}_0)
\end{aligned}
\tag{17}
$$

So

$$
D\left(\mathcal{N}(\boldsymbol{m}_1,\boldsymbol{\Sigma}_1)\,\|\,\mathcal{N}(\boldsymbol{m}_0,\boldsymbol{\Sigma}_0)\right) = \frac{1}{2}\left[\log\frac{|\boldsymbol{\Sigma}_0|}{|\boldsymbol{\Sigma}_1|}-n+\operatorname{tr}\left(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_0^{-1}\right)+(\boldsymbol{m}_1-\boldsymbol{m}_0)^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{m}_1-\boldsymbol{m}_0)\right]
\tag{18}
$$

(a) $D\left(\mathcal{N}(\boldsymbol{m}_1,\boldsymbol{\Sigma}_1)\,\|\,\mathcal{N}(\boldsymbol{m}_0,\boldsymbol{\Sigma}_0)\right) < \infty$ when $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_0$ is non-singular.

(b) With $\boldsymbol{m}_0 = 0, \boldsymbol{\Sigma}_0 = \boldsymbol{I}_n$ in the equation [18],

$$
D\left(\mathcal{N}(\boldsymbol{m}_1,\boldsymbol{\Sigma}_1)\,\|\,\mathcal{N}(\mathbf{0},\boldsymbol{I}_n)\right) = \frac{1}{2}\left[-\log|\boldsymbol{\Sigma}_1|-n+\operatorname{tr}(\boldsymbol{\Sigma}_1)+\boldsymbol{m}_1^T\boldsymbol{m}_1\right]
\tag{19}
$$

(c) As the equation [18] shows,

$$
D\left(\mathcal{N}(\boldsymbol{m}_1,\boldsymbol{\Sigma}_1)\,\|\,\mathcal{N}(\boldsymbol{m}_0,\boldsymbol{\Sigma}_0)\right) = \frac{1}{2}\left[\log\frac{|\boldsymbol{\Sigma}_0|}{|\boldsymbol{\Sigma}_1|}-n+\operatorname{tr}\left(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_0^{-1}\right)+(\boldsymbol{m}_1-\boldsymbol{m}_0)^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{m}_1-\boldsymbol{m}_0)\right]
\tag{20}
$$

2.6. (a) Actually, it is called Karamata's Inequality, and the proof is given as below.

Define

$$
c_i = \begin{cases} \dfrac{f(P_i)-f(Q_i)}{P_i-Q_i}, & \text{if } P_i \neq Q_i \\ f'(P_i), & \text{if } P_i = Q_i \end{cases}
\tag{21}
$$

Because f is a convex function, $\frac{f(x)-f(y)}{x-y}$ is monotonically non-decreasing for both x and y, we have $c_i \geqslant c_{i+1}$

Define $A_0 = B_0 = 0$, $A_n = \sum_{i=1}^n P_i, B_n = \sum_{i=1}^n Q_i$, so $A_n \leqslant B_n, 1 \leqslant n \leqslant k$, and $A_k = B_k$.

Consider

$$\begin{aligned}
\sum_{i=1}^{k}\left(f\left(P_i\right)-f\left(Q_i\right)\right) &= \sum_{i=1}^{n} c_i\left(P_i-Q_i\right) \\
&= \sum_{i=1}^{n} c_i(\underbrace{A_i-A_{i-1}}_{=P_i}-(\underbrace{B_i-B_{i-1}}_{=Q_i})) \\
&= \sum_{i=1}^{n} c_i\left(A_i-B_i\right)-\sum_{i=1}^{n} c_i\left(A_{i-1}-B_{i-1}\right) \\
&= c_n(\underbrace{A_n-B_n}_{=0})+\sum_{i=1}^{n-1}(\underbrace{c_i-c_{i+1}}_{\geqslant 0})(\underbrace{A_i-B_i}_{\leqslant 0})-c_1(\underbrace{A_0-B_0}_{=0}) \\
&\leqslant 0
\end{aligned} \tag{22}$$

Therefore,

$$\sum_{i=1}^{k} f\left(P_i\right) \leq \sum_{i=1}^{k} f\left(Q_i\right) \tag{23}$$

(b) Let $f(x)=x\log x$ is convex, using 23 and $H(P)=-\sum_{i=1}^{k} f\left(P_i\right)$, so

$$H(P) \geqslant H(Q) \tag{24}$$

2.7. *Total Correlation.*

(a) Part 1

$$\begin{aligned}
C\left(X_1,\ldots,X_n\right) &\triangleq D\left(P_{X^n} \| \prod_{i=1}^{n} P_{X_i}\right) \\
&= \mathbb{E}_{X^n}[\log\left(\frac{P_{X^n}}{\prod_{i=1}^{n} P_{X_i}}\right)] \\
&= \mathbb{E}_{X^n}[\log P_{X^n}]-\sum_{i=1}^{n} \mathbb{E}_{X^n}[\log P_{X_i}] \\
&= -H(X^n)-\sum_{i=1}^{n} \mathbb{E}_{X^i}[\log P_{X_i}] \\
&= \sum_{i=1}^{n} H(X_i)-H(X^n)
\end{aligned} \tag{25}$$

Part 2

Because $I(X^i;X_{i+1})=H(X^i)+H(X_{i+1})-H(X^i,X_{i+1})=$

$H(X^i) + H(X_{i+1}) - H(X^{i+1})$

$$\begin{aligned}\sum_{i=1}^{n-1} I(X^i; X_{i+1}) &= \sum_{i=1}^{n-1} H(X^i) + H(X_{i+1}) - H(X^{i+1}) \\ &= \sum_{i=1}^{n} H(X_i) - H(X^n) = C(X_1, \ldots, X_n)\end{aligned} \tag{26}$$

(b) $C(X_1, \ldots, X_n) \triangleq D\left(P_{X^n} \| \prod_{i=1}^n P_{X_i}\right) = 0$ when $P_{X^n} = \prod_{i=1}^n P_{X_i}$, that is, $X_1, X_2, \cdots, X_n$ are independent.

2.8. *Divergence of order statistics.*

(a) The joint distribution of order statistics is,

$$f_{X_{(1)},\ldots,X_{(n)}}(y_1, \ldots, y_n) = n! \prod_{i=1}^n f(y_i), \quad y_1 < y_2 < \cdots < y_n \tag{27}$$

Thus,

$$\begin{aligned}D\left(P_{X_{(1)}\ldots X_{(n)}} \| Q_{\,X_{(1)}\ldots X_{(n)})}\right) &= \mathbb{E}_{P_{X_{(1)}\cdots X_{(n)}}}\left[\log\left(\frac{P_{X_{(1)},\ldots,X_{(n)}}}{Q_{X_{(1)},\ldots,X_{(n)}}}\right)\right] \\ &= \mathbb{E}_{P_{X_{(1)}\cdots X_{(n)}}}\left[\log\left(\frac{\prod_{i=1}^n P(y_i)}{\prod_{i=1}^n Q(y_i)}\right)\right] \\ &= \sum_{i=1}^n \mathbb{E}_{P_{X_{(1)}\cdots X_{(n)}}}\left[\log\left(\frac{P(y_i)}{Q(y_i)}\right)\right] \\ &= \sum_{i=1}^n \mathbb{E}_{P_{X_1\cdots X_n}}\left[\log\left(\frac{P(x_i)}{Q(x_i)}\right)\right] \\ &= \sum_{i=1}^n \mathbb{E}_{P_{X_i}}\left[\log\left(\frac{P(x_i)}{Q(x_i)}\right)\right] \\ &= nD(P\|Q)\end{aligned} \tag{28}$$

(b) For $X \sim \text{Binom}(n, p)$, the p.m.f. is

$$P(X = i) = \binom{n}{i} p^i (1-p)^{n-i} \tag{29}$$

Thus,

$$\begin{aligned}D(\text{Binom}(n,p) \| \text{Binom}(n,q)) &= \sum_{i=0}^n \binom{n}{i} p^i (1-p)^{n-i} \log\left(\left(\frac{p(1-q)}{q(1-p)}\right)^i \left(\frac{1-p}{1-q}\right)^n\right) \\ &= n\log\left(\frac{1-p}{1-q}\right) + \sum_{i=0}^n i\binom{n}{i} p^i (1-p)^{n-i} \log\left(\frac{p(1-q)}{q(1-p)}\right) \\ &= n\log\left(\frac{1-p}{1-q}\right) + np\log\left(\frac{p(1-q)}{q(1-p)}\right)\end{aligned} \tag{30}$$

Let $n = 1$,

$$D(\mathrm{Bern}(p) \| \mathrm{Bern}(q)) = \log\left(\frac{1-p}{1-q}\right) + p\log\left(\frac{p(1-q)}{q(1-p)}\right) \tag{31}$$

So

$$D(\mathrm{Binom}(n,p) \| \mathrm{Binom}(n,q)) = nD(\mathrm{Bern}(p) \| \mathrm{Bern}(q)) \tag{32}$$