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

Homework 6

HANMO CHEN December 15, 2020

- **Acknowledgments:** For L_2-norm I refer to the wikipedia `https://en.wikipedia.org/wiki/Matrix_norm`. For 6.4, I refer to `https://ieeexplore.ieee.org/document/8849720`[1]
- **Collaborators:** I finish this homework all by myself.
- *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

6.1. (a)

$$p_{\mathsf{y}_1\mathsf{y}_2}(y_1, y_2; x) = 1, x \leqslant y_1, y_2 \leqslant x+1 \tag{1}$$

And $p(\mathsf{s} = \max(\mathsf{y}_1, \mathsf{y}_2) \leqslant s)) = (s-x)^2, x \leqslant s \leqslant x+1$, so

$$p_{\mathsf{s}}(s; x) = 2(s-x), x \leqslant s \leqslant x+1 \tag{2}$$

$$\frac{p_{\mathsf{y}_1\mathsf{y}_2}(y_1, y_2; x)}{p_{\mathsf{s}}(s; x)} = \frac{1}{2(s-x)}, x \leqslant y_1, y_2, s \leqslant x+1 \tag{3}$$

So s is not a sufficient statistic for $p_{\mathsf{y}_1\mathsf{y}_2}(y_1, y_2; x)$.

(b) $p_{\mathsf{r}}(r; x) = p_{\mathsf{r}}(-r; x)$ due to symmetry. Suppose $r \geqslant 0$,

$$p_{\mathsf{r}}(r; x) = \int_x^{x+1} p_{\mathsf{y}_1}(y+r; x) p_{\mathsf{y}_2}(y; x) dy = 1 - r \tag{4}$$

So $p_{\mathsf{r}}(r; x) = 1 - |r|$ does not depend on x, r is an ancillary statistic for $p_{\mathsf{y}_1\mathsf{y}_2}(y_1, y_2; x)$.

(c) Yes. Because we can recover $\mathsf{y}_1, \mathsf{y}_2$, with s, r.

$$(\mathsf{y}_1, \mathsf{y}_2) = \begin{cases} (\mathsf{s}, \mathsf{s} - \mathsf{r}), & r \geqslant 0 \\ (\mathsf{s} + \mathsf{r}, \mathsf{s}), & r < 0 \end{cases} \tag{5}$$

So $\mathbf{u} = [\mathsf{s}, \mathsf{r}]^T$ is a sufficient statistic for $p_{\mathsf{y}_1\mathsf{y}_2}(y_1, y_2; x)$ and

[1] S. Huang, X. Xu, L. Zheng and G. W. Wornell, "An Information Theoretic Interpretation to Deep Neural Networks," 2019 IEEE International Symposium on Information Theory (ISIT), Paris, France, 2019, pp. 1984-1988, doi: 10.1109/ISIT.2019.8849720.

6.2. Let $\mathsf{x} = \mathsf{y}$ be randomly chosen from $\{0, 1, 2\}$ with equal probability $\frac{1}{3}$ and $\xi(\mathsf{x}) = \mathsf{x}, \eta(\mathsf{y}) = \mathsf{y}^2$. $\rho(\mathsf{x}, \mathsf{y}) = \rho(\mathsf{x}, \mathsf{x}) = 1$.

$$\rho(\xi(\mathsf{x}), \eta(\mathsf{y})) = \rho(\mathsf{x}, \mathsf{x}^2) = \frac{\mathbb{E}[\mathsf{x}^3] - \mathbb{E}[\mathsf{x}^2]\,\mathbb{E}[\mathsf{x}]}{\sqrt{\operatorname{var}(\mathsf{x})\operatorname{var}(\mathsf{x}^2)}} = \sqrt{\frac{12}{13}} \neq \rho(\mathsf{x}, \mathsf{y}) \tag{6}$$

6.3. (a) i.

$$\begin{aligned} \boldsymbol{\psi}^T \mathbf{B} \boldsymbol{\phi} &= \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} g(y) \sqrt{P_{\mathsf{y}}(y)} \frac{P_{\mathsf{xy}}(x, y)}{\sqrt{P_{\mathsf{x}}(x) P_{\mathsf{y}}(y)}} f(x) \sqrt{P_{\mathsf{x}}(x)} \\ &= \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} f(x) g(y) P_{\mathsf{xy}}(x, y) \\ &= \mathbb{E}[f(x) g(y)] \end{aligned} \tag{7}$$

ii.

$$\begin{aligned} \mathbf{B}\boldsymbol{\phi}(y) &= \sum_{x \in \mathcal{X}} \frac{P_{\mathsf{xy}}(x, y)}{\sqrt{P_{\mathsf{y}}(y)}} f(x) \\ &= \sqrt{P_{\mathsf{y}}(y)} \sum_{x \in \mathcal{X}} P_{\mathsf{x}|\mathsf{y}}(x|y) f(x) \\ &= \sqrt{P_{\mathsf{y}}(y)}\, \mathbb{E}[f(\mathsf{x})|y] \end{aligned} \tag{8}$$

So $\mathbf{B}\boldsymbol{\phi} \leftrightarrow \mathbb{E}[f(\mathsf{x})|\mathsf{y}]$

iii.

$$\begin{aligned} \mathbf{B}^T \boldsymbol{\psi}(x) &= \sum_{y \in \mathcal{Y}} \frac{P_{\mathsf{xy}}(x, y)}{\sqrt{P_{\mathsf{x}}(x)}} g(y) \\ &= \sqrt{P_{\mathsf{x}}(x)} \sum_{y \in \mathcal{Y}} P_{\mathsf{y}|\mathsf{x}}(y|x) g(y) \\ &= \sqrt{P_{\mathsf{x}}(x)}\, \mathbb{E}[g(\mathsf{y})|x] \end{aligned} \tag{9}$$

So $\mathbf{B}^T \boldsymbol{\psi} \leftrightarrow \mathbb{E}[g(\mathsf{y})|\mathsf{x}]$

(b)

$$\mathbf{B}\boldsymbol{\phi}_1(y) = \sum_{x \in \mathcal{X}} \frac{P_{\mathsf{xy}}(x, y)}{\sqrt{P_{\mathsf{y}}(y)}} = \sqrt{P_{\mathsf{y}}(y)} \tag{10}$$

$$\mathbf{B}^T \boldsymbol{\psi}_1(x) = \sum_{y \in \mathcal{Y}} \frac{P_{\mathsf{xy}}(x, y)}{\sqrt{P_{\mathsf{x}}(x)}} = \sqrt{P_{\mathsf{x}}(x)} \tag{11}$$

So $\mathbf{B}\boldsymbol{\phi}_1 = \boldsymbol{\psi}_1, \mathbf{B}^T \boldsymbol{\psi}_1 = \boldsymbol{\phi}_1$. Its means that if we let $f(x) = g(y) = 1$ in (a), we have $\mathbb{E}[1|\mathsf{y}] = \mathbb{E}[1|\mathsf{x}] = 1$

(c) The L_2-norm of $\mathbf{B}$ is equivalent to

$$\|\mathbf{B}\|_2 = \sup \|\mathbf{B}\boldsymbol{x}\|_2, \text{subject to} \|\boldsymbol{x}\|_2 = 1 \tag{12}$$

To maximize $\|\mathbf{B}\boldsymbol{x}\|_2$, every component in $\boldsymbol{x}$ must $\geqslant 0$. Thus for every $\mathbf{X} \in \mathbb{R}^{|\mathcal{X}|}$, there is a correspond distribution of $X \in \mathcal{X}$. So $\mathbf{X}$ can be expressed as

$$\boldsymbol{x} = \left[\sqrt{P_{\mathsf{x}'}(1)}, \sqrt{P_{\mathsf{x}'}(2)}, \cdots, \sqrt{P_{\mathsf{x}'}(|\mathcal{X}|)}\right]^T \tag{13}$$

Thus

$$\begin{aligned}
\|\mathbf{B}\boldsymbol{x}\|_2^2 &= \sum_{y\in\mathcal{Y}} \left(\sum_{x\in\mathcal{X}} \frac{P_{\mathsf{xy}}(x,y)}{\sqrt{P_{\mathsf{x}}(x)P_{\mathsf{y}}(y)}} \sqrt{P_{\mathsf{x}'}(x)}\right)^2 \\
&= \sum_{y\in\mathcal{Y}} P_{\mathsf{y}}(y) \left(\sum_{x\in\mathcal{X}} \frac{P_{\mathsf{x}|\mathsf{y}}(x|y)}{\sqrt{P_{\mathsf{x}}(x)}} \sqrt{P_{\mathsf{x}'}(x)}\right)^2 \\
&= \sum_{y\in\mathcal{Y}} P_{\mathsf{y}}(y)\, \mathbb{E}^2 \left[\frac{\sqrt{P_{\mathsf{x}'}(x)}}{\sqrt{P_{\mathsf{x}}(x)}} |y\right] \\
&\leqslant \sum_{y\in\mathcal{Y}} P_{\mathsf{y}}(y)\, \mathbb{E} \left[\frac{P_{\mathsf{x}'}(x)}{P_{\mathsf{x}}(x)} |y\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{P_{\mathsf{x}'}(x)}{P_{\mathsf{x}}(x)} |y\right]\right] \\
&= \mathbb{E}[\frac{P_{\mathsf{x}'}(x)}{P_{\mathsf{x}}(x)}] = 1
\end{aligned} \tag{14}$$

And the "=" holds when $\boldsymbol{x} = \boldsymbol{\phi}_1$. So $\|\mathbf{B}\|_2 = 1$

6.4. (a) The empirical mean $\frac{1}{n}\sum_{i=1}^n \log Q_{\mathsf{y}|\mathsf{x}}(y_i|x_i)$ is the realization of expectation $\mathbb{E}_{P_{\mathsf{xy}}}[\log Q_{\mathsf{y}|\mathsf{x}}(x,y)]$.

$$\begin{aligned}
(\boldsymbol{g}^*, b^*) &= \arg\max_{\boldsymbol{g},\boldsymbol{b}} D(P_{\mathsf{xy}}\|P_{\mathsf{x}}Q_{\mathsf{y}|\mathsf{x}}) \\
&= \arg\max_{\boldsymbol{g},\boldsymbol{b}} \mathbb{E}_{P_{\mathsf{xy}}}\left[\log \frac{P_{\mathsf{xy}}(x,y)}{P_{\mathsf{x}}(x)Q_{\mathsf{y}|\mathsf{x}}(y|x)}\right] \\
&= \arg\min_{\boldsymbol{g},\boldsymbol{b}} \mathbb{E}_{P_{\mathsf{xy}}}[\log Q_{\mathsf{y}|\mathsf{x}}(y|x)] \\
&= \arg\min_{\boldsymbol{g},\boldsymbol{b}} \frac{1}{n}\sum_{i=1}^n \log Q_{\mathsf{y}|\mathsf{x}}(y_i|x_i)
\end{aligned} \tag{15}$$

(b) First, we will prove the fisrt order approximation of K-L divergence, i.e. if $P_1(x) - P_2(x) = O(\varepsilon)$,

$$
\begin{aligned}
D(P_1(x)\|P_2(x)) &= -\sum_{x\in\mathcal{X}} P_1(x)\log\left(\frac{P_1(x)+P_2(x)-P_1(x)}{P_1(x)}\right)\\
&= -\sum_{x\in\mathcal{X}} P_1(x)\left(\frac{P_2(x)-P_1(x)}{P_1(x)}-\frac{1}{2}\frac{(P_2(x)-P_1(x))^2}{(P_1(x))^2}+O(\epsilon^3)\right)\\
&\approx \frac{1}{2}\sum_{x\in\mathcal{X}}\frac{(P_2(x)-P_1(x))^2}{P_1(x)} = \frac{1}{2}\sum_{x\in\mathcal{X}}\frac{(P_2(x)-P_1(x))^2}{P_2(x)}\frac{P_2(x)}{P_1(x)}\\
&= \frac{1}{2}\sum_{x\in\mathcal{X}}\frac{(P_2(x)-P_1(x))^2}{P_2(x)}(1-\frac{O(\epsilon)}{P_1(x)})\\
&\approx \frac{1}{2}\sum_{x\in\mathcal{X}}\frac{(P_2(x)-P_1(x))^2}{P_2(x)}
\end{aligned}
\tag{16}
$$

Then we just need to prove that $P_{\mathsf{xy}}(x,y)-P_{\mathsf{x}}(x)Q_{\mathsf{y}|\mathsf{x}}(y)=O(\epsilon)$.
Because $\boldsymbol{f}^T(x)\boldsymbol{g}^*(y)+d^*(y)=O(\varepsilon)$,

$$
P_{\mathsf{y}}(y)e^{\boldsymbol{f}^T(x)\boldsymbol{g}^*(y)+d^*(y)} = P_{\mathsf{y}}(y)(1+\boldsymbol{f}^T(x)\boldsymbol{g}^*(y)+d^*(y)+O(\varepsilon^2)) \tag{17}
$$

$$
\sum_{y\in\mathcal{Y}} P_{\mathsf{y}}(y)(1+\boldsymbol{f}^T(x)\boldsymbol{g}^*(y)+d^*(y)+O(\varepsilon^2)) = 1+\mathbb{E}_{P_{\mathsf{y}}}[\boldsymbol{f}^T(x)\boldsymbol{g}^*(y)]+\mathbb{E}_{P_{\mathsf{y}}}[d^*(y)]+O(\varepsilon^2) \tag{18}
$$

Without loss of generality, we can assume
$\mathbb{E}_{P_{\mathsf{y}}}[\boldsymbol{g}^*(y)]=\mathbb{E}_{P_{\mathsf{y}}}[d^*(y)]=0$.

$$
\begin{aligned}
Q_{\mathsf{y}|\mathsf{x}}[(x,y)] &= \frac{P_{\mathsf{y}}(y)(1+\boldsymbol{f}^T(x)\boldsymbol{g}^*(y)+d^*(y)+O(\varepsilon^2))}{\sum_{y'\in\mathcal{Y}} P_{\mathsf{y}}(y')(1+\boldsymbol{f}^T(x)\boldsymbol{g}^*(y')+d^*(y')+O(\varepsilon^2))}\\
&= \frac{P_{\mathsf{y}}(y)\left(1+\boldsymbol{f}^T(x)\boldsymbol{g}^*(y)+d^*(y)+O(\varepsilon^2)\right)}{1+O(\varepsilon^2)}\\
&= P_{\mathsf{y}}(y)\left(1+\boldsymbol{f}^T(x)\boldsymbol{g}^*(y)+d^*(y)+O(\varepsilon^2)\right)(1-O(\varepsilon^2))\\
&= P_{\mathsf{y}}(y)\left(1+\boldsymbol{f}^T(x)\boldsymbol{g}^*(y)+d^*(y)\right)+O(\varepsilon^2)
\end{aligned}
\tag{19}
$$

Thus,

$$
P_{\mathsf{xy}}(x,y)-P_{\mathsf{x}}(x)Q_{\mathsf{y}|\mathsf{x}}(y) = P_{\mathsf{x}}(x)P_{\mathsf{y}}(y)+O(\varepsilon)-P_{\mathsf{x}}(x)P_{\mathsf{y}}(y)(1+O(\varepsilon)) = O(\varepsilon) \tag{20}
$$

So we have,

$$
D(P_{\mathsf{xy}}(x,y)\|P_{\mathsf{x}}(x)Q_{\mathsf{y}|\mathsf{x}}(y)) \approx \frac{1}{2}\sum_{x\in\mathcal{X},y\in\mathcal{Y}}\frac{(P_{\mathsf{xy}}(x,y)-P_{\mathsf{x}}(x)Q_{\mathsf{y}|\mathsf{x}}(y))^2}{P_{\mathsf{x}}(x)Q_{\mathsf{y}|\mathsf{x}}(y)} \tag{21}
$$