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

Homework 5

HANMO CHEN December 1, 2020

- **Acknowledgments:** For Problem 1.2(b), I refer to some answer on StackExchange `https://math.stackexchange.com/a/108303`
- **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

5.1. *Cramer-Rao inequality with a bias term.*

Proof. The bias $b(x) = \mathbb{E}[\hat{x}(\mathsf{y})] - x = \int_{\mathbb{R}} \hat{x}(y) f(y;x) dy - x$. So $b'(x) = \int_{\mathbb{R}} \hat{x}(y) \frac{\partial f(y;x)}{\partial x} dy - 1$

Notice that

$$\begin{aligned} \mathbb{E}\left[(\hat{x}(\mathsf{y}) - x)^2\right] - b^2(x) &= \mathbb{E}\left[(\hat{x}(\mathsf{y}) - x)^2\right] - (\mathbb{E}[\hat{x}(\mathsf{y}) - x])^2 \\ &= \mathrm{var}(\hat{x}(\mathsf{y}) - x) = \mathrm{var}(\hat{x}(\mathsf{y})) \end{aligned} \tag{1}$$

So the origincal inequality is equivalent to the following one,

$$\mathrm{var}(\hat{x}(\mathsf{y})) \geqslant \frac{[1 + b'(x)]^2}{J_{\mathsf{y}}(x)} \tag{2}$$

where $J_{\mathsf{y}}(x) = \mathbb{E}\left[\left(\frac{\partial}{\partial x} \ln f(y;x)\right)^2\right]$.

$$1 + b'(x) = \int_{\mathbb{R}} \hat{x}(y) \frac{\partial f(y;x)}{\partial x} dy = \int_{\mathbb{R}} \hat{x}(y) \frac{\frac{\partial f(y;x)}{\partial x}}{f(y;x)} f(y;x) dy \tag{3}$$

Notice that

$$\frac{\partial}{\partial x} \ln f(y;x) = \frac{\frac{\partial f(y;x)}{\partial x}}{f(y;x)} \tag{4}$$

So

$$1 + b'(x) = \mathbb{E}\left[\hat{x}(y) \frac{\partial}{\partial x} \ln f(y;x)\right] \tag{5}$$

And according to the regularity condition,

$$\mathbb{E}\left[\frac{\partial}{\partial x}\ln f(y;x)\right] = 0 \tag{6}$$

Thus,

$$1 + b'(x) = \mathbb{E}\left[(\hat{x}(y) - \mathbb{E}[\hat{x}(y)])\frac{\partial}{\partial x}\ln f(y;x)\right] \tag{7}$$

Using the Cauchy-Schwarz inequality,

$$\begin{aligned}(1 + b'(x))^2 &= \left(\mathbb{E}\left[(\hat{x}(y) - \mathbb{E}[\hat{x}(y)])\frac{\partial}{\partial x}\ln f(y;x)\right]\right)^2 \\ &\leqslant \mathbb{E}\left[(\hat{x}(y) - \mathbb{E}[\hat{x}(y)])^2\right]\mathbb{E}\left[\left(\frac{\partial}{\partial x}\ln f(y;x)\right)^2\right]\end{aligned} \tag{8}$$

that is,

$$\mathrm{var}(\hat{x}(y)) \geqslant \frac{[1 + b'(x)]^2}{J_{\mathsf{y}}(x)} \tag{9}$$

$\square$

5.2. (a) Suppose an unbiased estimator $f(y)$ for x exists, so for all x >0,

$$\int_0^{\frac{1}{x}} xf(y)dy = x \Longrightarrow \int_0^{\frac{1}{x}} f(y)dy = 1 \tag{10}$$

That is, $\forall a > 0$

$$\int_0^a f(y)dy = 1 \tag{11}$$

$\forall b > a > 0$,

$$\int_a^b f(y)dy = 0 \tag{12}$$

$f(y) = 0, \forall y > 0$, which is not an unbiased estimator. So there is no unbiased estimator for x.

(b) First we try to derive an unbiased estimator, $\forall x > 0$

$$\int_0^x \frac{1}{x}f(y)dy = x \Longrightarrow \int_0^x f(y)dy = x^2 \tag{13}$$

Obviously $f(y) = 2y$ is an unbiased estimator. It is easy to show that y is a complete sufficient statistics of x, so $\hat{x}(y) = 2y$ is a minimum-variance unbiased estimator for x.

5.3. (a) The distribution function is

$$f(\underline{y};x) = \begin{cases} \frac{1}{2\pi}\exp\left(-\frac{(y_1-x)^2}{2}-\frac{(y_2-x)^2}{2}\right), & x>0 \\ \frac{1}{2\sqrt{2}\pi}\exp\left(-\frac{(y_1-x)^2}{2}-\frac{(y_2-x)^2}{4}\right), & x<0 \end{cases} \tag{14}$$

$$\frac{\partial \ln f(\underline{y};x)}{\partial x} = \begin{cases} y_1+y_2-2x, & x>0 \\ y_1+\frac{y_2}{2}-\frac{3}{2}x, & x<0 \end{cases} \tag{15}$$

$$J_{\underline{\mathsf{y}}}(x) = \mathbb{E}\left[\left(\frac{\partial \ln f(\underline{y};x)}{\partial x}\right)^2\right] = \begin{cases} 2, & x>0 \\ \frac{3}{2}, & x<0 \end{cases} \tag{16}$$

(b) Consider the following estimators,

$$\begin{aligned} \hat{x}_1(\underline{\mathsf{y}}) &= \frac{1}{2}\mathsf{y}_1+\frac{1}{2}\mathsf{y}_2 \\ \hat{x}_2(\underline{\mathsf{y}}) &= \frac{2}{3}\mathsf{y}_1+\frac{1}{3}\mathsf{y}_2 \end{aligned} \tag{17}$$

It is easy to show that $\hat{x}_1(\underline{\mathsf{y}}), \hat{x}_2(\underline{\mathsf{y}})$ are unbiased estimators, and

$$\begin{aligned} \mathrm{var}(\hat{x}_1(\underline{\mathsf{y}})) &= \begin{cases} \frac{1}{2}, x>0 \\ \frac{1}{4}, x<0 \end{cases} \\ \mathrm{var}(\hat{x}_2(\underline{\mathsf{y}})) &= \begin{cases} \frac{5}{9}, x>0 \\ \frac{2}{3}, x<0 \end{cases} \end{aligned} \tag{18}$$

For $x>0$, $\hat{x}_1(\underline{\mathsf{y}})$ achieves the Cramer-Rao Lower Bound and $x<0$, $\hat{x}_2(\underline{\mathsf{y}})$ achieves the Cramer-Rao Lower Bound. So there is no minimal-variance unbiased estimator for x.

5.4. (a)

$$P_{\underline{\mathsf{y}}}(\underline{y};x) = x^{y_1+y_2}(1-x)^{2-y_1-y_2} = (1-x)^2\left(\frac{x}{1-x}\right)^{t(\underline{y})} = a(t(\underline{y}),x)b(x) \tag{19}$$

So $t(\underline{y}) = y_1+y_2$ is a sufficient statistics for x.

(b)

$$\mathrm{MSE}_{\hat{x}}(x) = \mathbb{E}\left[(x-\hat{x}(\underline{y}))^2\right] = x(1-x) \tag{20}$$

(c) i. Because $t(\underline{y})$ is a sufficient statistics, $P_{\underline{\mathsf{y}}}(\underline{y};x) = P_{\underline{\mathsf{y}}|T}(\underline{y}|t)P_T(t;x)$. So $P_{\underline{\mathsf{y}}|T}(\underline{y}|t)$ is independent of x. And $x'(t) = \mathbb{E}[\hat{x}(\underline{\mathsf{y}})|\mathsf{t}=t]$ does not depend on x.

$$x'(t) = \mathbb{E}[y_1|\mathsf{t}=t] = \begin{cases} 0, t=0 \\ \frac{1}{2}, t=1 \\ 1, t=2 \end{cases} \tag{21}$$

ii. $x'(t) = \frac{1}{2}t = \frac{y_1+y_2}{2}$.

$$\mathrm{MSE}_{\hat{x}'}(x) = \mathbb{E}\left[(x - \frac{y_1+y_2}{2})^2\right] = \frac{2}{2}x(1-x) \tag{22}$$

So $MSE_{\hat{x}'}(x) = \frac{1}{2}MSE_{\hat{x}}(x)$

(d) i. Because $t(\underline{y})$ is a sufficient statistics,
$P_{\underline{\mathsf{y}}}(\underline{y};x) = P_{\underline{\mathsf{y}}|T}(\underline{y}|t)P_T(t;x)$. So $P_{\underline{\mathsf{y}}|T}(\underline{y}|t)$ is independent of x.
And $x'(t) = \mathbb{E}[\hat{x}(\underline{y})|\mathsf{t}=t]$ does not depend on x.

ii. *(Rao-Blackwell Theorem)* Because the cost function $C(x,\hat{x})$ is convex in $\hat{x}$, using Jensen's inequality, $\forall t$

$$C(x,\hat{x}'(t)) = C(x,\mathbb{E}[\hat{x}(\underline{\mathsf{y}})|\mathsf{t}=t]) \leqslant \mathbb{E}\left[C(x,\hat{x}(\underline{\mathsf{y}})) \mid \mathsf{t}=t\right] \tag{23}$$

Take expections of t in the inequality above,

$$\mathbb{E}[C(x,\hat{x}'(\mathsf{t}))] \leqslant \mathbb{E}\left[C(x,\hat{x}(\underline{\mathsf{y}}))\right] \tag{24}$$

5.5. First we prove it is a sufficient statistics.

$$p_{\mathsf{y}}(y;x) = \frac{p_{\mathsf{y}}(y;H_0)}{p_{\mathsf{y}}(y;H_1)}\frac{p_{\mathsf{y}}(y;x)}{p_{\mathsf{y}}(y;H_0)}p_{\mathsf{y}}(y;H_1) = a(t(y),x)b(y) \tag{25}$$

$$\begin{aligned} a(t(y),x) &= \frac{p_{\mathsf{y}}(y;H_0)}{p_{\mathsf{y}}(y;H_1)}\frac{p_{\mathsf{y}}(y;x)}{p_{\mathsf{y}}(y;H_0)} = \begin{cases} 1, x = H_1 \\ t(y), x = H_0 \end{cases} \\ b(y) &= p_{\mathsf{y}}(y;H_1) \end{aligned} \tag{26}$$

So $t(y)$ is a sufficient statistics.

Notice that the distribution function belongs to an exponential family,

$$\begin{aligned} p_{\mathsf{y}}(y;x) &= p_{\mathsf{y}}(y;H_1)\exp\left(\mathbb{1}(x=H_0)\log\frac{p_{\mathsf{y}}(y;H_0)}{p_{\mathsf{y}}(y;H_1)}\right) \\ &= h(y)\exp\left(w(x)t(y)\right) \end{aligned} \tag{27}$$

Thus $t(y)$ is a complete statistics.

So $t(y)$ is a complete sufficient statistics

5.6.

$$\hat{x}_{MAP}(y) = \arg\max_a p_{\mathsf{x}|\mathsf{y}}(a|y) = \arg\max_a p_{\mathsf{x},\mathsf{y}}(a,y) = \arg\max_a p_{\mathsf{y}|\mathsf{x}}(y|a)p_{\mathsf{x}}(a) \tag{28}$$

Suppose z is the complete data,

$$L(x) = \log p_{\mathsf{y}|\mathsf{x}}(y|x)p_{\mathsf{x}}(x) = \log p_{\mathsf{z}}(z|x) - \log p_{\mathsf{z}|\mathsf{y},\mathsf{x}}(z|y,x) + \log p_{\mathsf{x}}(x) \tag{29}$$

Take expections of both sides over $p_{\mathsf{z}|\mathsf{y},\mathsf{x}}(z|y,x')$

$$LHS = \log p_{\mathsf{y}|\mathsf{x}}(y|x)p_{\mathsf{x}}(x) \tag{30}$$

$$RHS = \sum_{z'} p_{\mathsf{z}|\mathsf{y},\mathsf{x}}(z'|y,x') \log p_{\mathsf{z}}(z'|x) - \sum_{z'} p_{\mathsf{z}|\mathsf{y},\mathsf{x}}(z'|y,x') \log p_{\mathsf{z}|\mathsf{y},\mathsf{x}}(z'|y,x) + \log p_{\mathsf{x}}(x) \tag{31}$$

Denote $U(x,x') = \sum_{z'} p_{\mathsf{z}|\mathsf{y},\mathsf{x}}(z'|y,x') \log p_{\mathsf{z}}(z'|x) + \log p_{\mathsf{x}}(x)$ and $V(x,x') = -\sum_{z'} p_{\mathsf{z}|\mathsf{y},\mathsf{x}}(z'|y,x') \log p_{\mathsf{z}|\mathsf{y},\mathsf{x}}(z'|y,x)$ so

$$\log p_{\mathsf{y}|\mathsf{x}}(y|x)p_{\mathsf{x}}(x) = U(x,x') + V(x,x') \tag{32}$$

By the same method in lecture, we have

$$V(x,x') - V(x',x') = \sum_{z} p_{\mathsf{z}|\mathsf{y},\mathsf{x}}(z|y,x') \log \frac{p_{\mathsf{z}|\mathsf{y},\mathsf{x}}(z|y,x')}{p_{\mathsf{z}|\mathsf{y},\mathsf{x}}(z|y,x)} = D(p_{\mathsf{z}|\mathsf{y},\mathsf{x}'} \| p_{\mathsf{z}|\mathsf{y},\mathsf{x}}) \geqslant 0 \tag{33}$$

So if we want to find $L(x) > L(x')$, we just need to make sure $U(x,x') > U(x',x')$.

The EM-MAP algorithm is as follows.

(a) Initialize: choose $x^{(0)}$

(b) Repeat until convergence:

   i. E-step: given previous $x^{(n)}$, compute

$$U(x,x^{(n)}) = \mathbb{E}_{p_{\mathsf{z}|\mathsf{y},\mathsf{x}}}\left[\log p_{\mathsf{z}}(z|x) \middle| \mathsf{y} = y, \mathsf{x} = x^{(n)}\right] + \log p_{\mathsf{x}}(x) \tag{34}$$

   ii. M-step: determine $x^{(n+1)}$,

$$x^{(n+1)} = \arg\max_{x} U(x,x^{(n)}) \tag{35}$$