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

Problem Set 5

Issued: Monday 30th November, 2020 **Due:** Monday 14th December, 2020

Notations: We use $\mathsf{x}, \mathsf{y}, \mathsf{w}$ and $\underline{\mathsf{x}}, \underline{\mathsf{y}}, \underline{\mathsf{w}}$ to denote random variables and random vectors.

5.1. *Cramer-Rao inequality with a bias term.* Let $\mathsf{y} \sim f(y; x)$ and let $\hat{x}(\mathsf{y})$ be an estimator for x. Let $b(x) = \mathbb{E}[\hat{x}(\mathsf{y})] - x$ be the bias of the estimator. Show that

$$\mathbb{E}\left[(\hat{x}(\mathsf{y}) - x)^2\right] \geq \frac{[1 + b'(x)]^2}{J_{\mathsf{y}}(x)} + b^2(x)$$

Solution:

$$\begin{aligned}
\mathbb{E}[\frac{\partial}{\partial x} \ln f(\mathsf{y}; x) \cdot \hat{x}(\mathsf{y})] &= \int \frac{\partial}{\partial x} f(y; x)\hat{x}(y)\mathrm{d}y \\
&= \frac{\partial}{\partial x} \mathbb{E}[\hat{x}(\mathsf{y})] \\
&= 1 + b'(x)
\end{aligned}$$

By the Cauchy-Schwarz inequality, we have

$$\mathrm{Cov}^2[e(\mathsf{y}), s(\mathsf{y}; x)] \leq \mathrm{Var}[e(\mathsf{y})]\mathrm{Var}[s(\mathsf{y}; x)].$$

It leads to

$$(1 + b'(x))^2 \leq J_{\mathsf{y}}(x)\, \mathbb{E}[(\hat{x}(\mathsf{y}) - \mathbb{E}[\hat{x}(\mathsf{y})])^2] = J_{\mathsf{y}}(x) \left(\mathbb{E}\left[(\hat{x}(\mathsf{y}) - x)^2\right] - b^2(x)\right)$$

That's the inequality.

5.2. (a) Let

$$p_{\mathsf{y}}(y; x) = \begin{cases} x & \text{if } 0 \leq y \leq 1/x \\ 0 & \text{otherwise} \end{cases}$$

for $x > 0$. Show that there exist no unbiased estimators $\hat{x}(\mathsf{y})$ for x. (Note that because only $x > 0$ are possible values, an unbiased estimator need only be unbiased for $x > 0$ rather than all x.)

(b) Suppose instead that

$$p_{\mathsf{y}}(y; x) = \begin{cases} 1/x & \text{if } 0 \leq y \leq x \\ 0 & \text{otherwise} \end{cases}$$

for $x > 0$. Does a minimum-variance unbiased estimator for x based on y exist? If your answer is yes, determine $\hat{x}_{\mathrm{MVU}}(\mathsf{y})$. If your answer is no, explain.

Solution:

(a) From

$$\int_0^{1/x} p_{\mathsf{y}}(y;x)\hat{x}(y)dy = x$$

we have

$$\int_0^{1/x} \hat{x}(y)dy = 1,$$

i.e.,

$$\int_0^{t} \hat{x}(y)dy = 1, \text{ for all } t > 0.$$

Therefore, we have $\hat{x}(y) = 0$ for all $y > 0$, and thus

$$\int_0^{t} \hat{x}(y)dy = 0,$$

which leads to a contradiction.

(b) It can be shown that $\hat{x}(y) = 2y$ is the only unbiased estimator, thus it is the MVU estimator.

5.3. Suppose, for $i = 1, 2$

$$\mathsf{y}_i = x + \mathsf{w}_i$$

where x is an unknown but non-zero constant, w_1 and w_2 are statistically independent, zero-mean Gaussian random variables with

$$\begin{aligned} \text{var}(\mathsf{w}_1) &= 1 \\ \text{var}(\mathsf{w}_2) &= \begin{cases} 1 & x > 0 \\ 2 & x < 0 \end{cases}. \end{aligned}$$

(a) Calculate the Cramér-Rao bound for unbiased estimators of x based on observation of

$$\underline{\mathsf{y}} = \begin{bmatrix} \mathsf{y}_1 \\ \mathsf{y}_2 \end{bmatrix}.$$

(b) Show that a minimum variance unbiased estimator $\hat{x}_{\text{MVU}}(\underline{\mathsf{y}})$ does not exist.

Hint: Consider the estimators

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

Solution:

(a) The Cramér-Rao bound is

$$\lambda_e(x) \geq \begin{cases} \frac{1}{2}, & x > 0 \\ \frac{2}{3}, & x < 0. \end{cases}$$

(b) It can be verified that both $\hat{x}_1(\underline{\mathsf{y}})$ and $\hat{x}_2(\underline{\mathsf{y}})$ are unbiased, and has error variances

$$\lambda_1(x) = \begin{cases} \frac{1}{2}, & x > 0 \\ \frac{3}{4}, & x < 0. \end{cases}$$

$$\lambda_2(x) = \begin{cases} \frac{5}{9}, & x > 0 \\ \frac{2}{3}, & x < 0. \end{cases}$$

Then, if a MVU exists, it must have a variance no greater than that of these two, i.e.,

$$\lambda(x) \leq \begin{cases} \frac{1}{2}, & x > 0 \\ \frac{2}{3}, & x < 0 \end{cases},$$

and thus must be efficient. However, the efficient estimator is

$$\hat{x}_{\text{eff}}(\underline{\mathsf{y}}) = \begin{cases} \hat{x}_1(\underline{\mathsf{y}}), & x > 0 \\ \hat{x}_2(\underline{\mathsf{y}}), & x < 0 \end{cases},$$

which depends on x and is not valid. Therefore, a MVU does not exist.

5.4. Let $\underline{\mathsf{y}} = [\mathsf{y}_1\ \mathsf{y}_2]^{\mathrm{T}}$ be a vector random variable whose components are i.i.d. Bernoulli random variables with parameter x, $0 < x < 1$, i.e., $\mathbb{P}(\mathsf{y}_i = 1) = x, i = 1, 2$.

(a) Show that $t(\underline{y}) = y_1 + y_2$ is a sufficient statistic.

(b) Let $\hat{x}(\underline{y}) = y_1$ be an estimator of the parameter x from the observation $\underline{y}$. Find $\mathrm{MSE}_{\hat{x}}(x)$, the mean-square error of this estimator.

(c) Let $\hat{x}'(t) = \mathbb{E}[\hat{x}(\underline{\mathsf{y}})|\mathsf{t} = t]$ be an estimator of the parameter x that uses the sufficient statistic t instead of the observations $\underline{y}$.

i. Show that $\hat{x}'(t)$ is a valid estimator, i.e., it does not depend on x.

ii. Show that $\mathrm{MSE}_{\hat{x}'}(x) = \gamma\, \mathrm{MSE}_{\hat{x}}(x)$ and find the constant γ.

(d) We now consider a generalization of this problem. Let $\underline{\mathsf{y}}$ be a random variable generated by a distribution $p_{\underline{\mathsf{y}}}(\cdot; x)$ and $\underline{t}(\underline{y})$ be a sufficient statistic. Let $\hat{x}(\underline{y})$ be an estimator of the parameter x based on the observation $\underline{y}$. We define an alternate estimator $\hat{x}'(\underline{t}) = \mathbb{E}[\hat{x}(\underline{\mathsf{y}})|\underline{\mathsf{t}} = \underline{t}]$.

i. Show that $\hat{x}'(\underline{t})$ is a valid estimator, i.e., it does not depend on x.

ii. Show that for any cost function $C(x, \hat{x})$ that is convex in $\hat{x}$, the following inequality holds:

$$\mathbb{E}[C(x, \hat{x}'(\underline{\mathsf{t}}))] \leq \mathbb{E}[C(x, \hat{x}(\underline{\mathsf{y}}))].$$

Solution:

(a) Note that $p_{\underline{\mathsf{y}}}(\underline{y}; x) = x^{y_1+y_2}(1-x)^{2-(y_1+y_2)} = x^{t(\underline{y})}(1-x)^{2-t(\underline{y})}$.

(b) $\mathrm{MSE}_{\hat{x}}(x) = x(1-x)$

(c) i. Easily verified by computing all possible values of $\hat{x}'(t)$.

ii. $\gamma = \frac{1}{2}$.

(d) i. Easily verified by the definition of sufficiency.

ii. By Jensen's inequality, we have

$$C(x, \hat{x}'(\underline{\mathsf{t}})) = C(x, \mathbb{E}[\hat{x}(\underline{\mathsf{y}})|\underline{\mathsf{t}}]) \leq \mathbb{E}[C(x, \hat{x}(\underline{\mathsf{y}}))|\underline{\mathsf{t}}],$$

and thus

$$\mathbb{E}[C(x, \hat{x}'(\underline{\mathsf{t}}))] \leq \mathbb{E}[\mathbb{E}[C(x, \hat{x}(\underline{\mathsf{y}}))|\underline{\mathsf{t}}]] = \mathbb{E}[C(x, \hat{x}(\underline{\mathsf{y}}))].$$

5.5. For a non-bayesian case $p_{\mathsf{y}}(y; x)$, we do a binary hypothesis testing where $x \in \{H_0, H_1\}$. Please prove that $t(y) = \frac{p_{\mathsf{y}}(y;H_1)}{p_{\mathsf{y}}(y;H_0)}$ is a complete sufficient statistic.

Solution: Firstly, in the lecture note, we have proved that $t(y)$ is a sufficient statistic.
Then we need to use the uniqueness of the Laplace transform, which means we should write down the parameterized density fucntion as an exponential family with $t(y)$. It is a normal way to explain the completeness, especially when we can not know the model. Please see the definition and properties of Laplace transform.
In this case $p_{\mathsf{y}}(y; x) = p_{\mathsf{y}}(y; H_0) \exp\left(\mathbb{1}\{x = H_1\} \log t(y)\right)$.

5.6. In class we developed the EM algorithm for maximum likelihood estimation (EM-ML). That is, we gave an iterative procedure to compute

$$\hat{x}_{ML}(y) = \arg\max_a p_{\mathsf{y}}(y; a).$$

and showed that the likelihood was non-decreasing with each iteration.
Please develop the EM-MAP algorithm for MAP estimation:

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

where the complete data z is an arbitrary random vector. (Please follow the procedures in the lecture note)

Solution: We start with the decomposition

$$p_{\mathsf{x}|\mathsf{y}}(x|y) = \frac{p_{\mathsf{xz}|\mathsf{y}}(xz|y)}{p_{\mathsf{z}|\mathsf{xy}}(z|xy)}$$

Next, we take the expectation with respect to the distribution $p_{\mathsf{z}|\mathsf{xy}}(z|xy)$

$$\log p_{\mathsf{x}|\mathsf{y}}(x|y) = \mathbb{E}[\log p_{\mathsf{xz}|\mathsf{y}}(xz|y)|\mathsf{x} = x', \mathsf{y} = y] - \mathbb{E}[\log p_{\mathsf{z}|\mathsf{xy}}(z|xy)|\mathsf{x} = x', \mathsf{y} = y],$$

which we write in a more compact form as

$$B_{\mathsf{y}}(x) = U_{\mathsf{y}}(x, x') - V_{\mathsf{y}}(x, x')$$

By the Gibbs' inequality, $V_{\mathsf{y}}(x, x') \leq V_{\mathsf{y}}(x', x')$. We compare $B_{\mathsf{y}}(x)$ at successive values:

$$B_{\mathsf{y}}(x) - B_{\mathsf{y}}(x') = [U_{\mathsf{y}}(x, x') - U_{\mathsf{y}}(x', x')] + [V_{\mathsf{y}}(x', x') - V_{\mathsf{y}}(x, x')]$$

Since the second term is non-negative, we can increase $B_{\mathsf{y}}(x)$ by increasing $U_{\mathsf{y}}(x, x')$. This justifies the following version of the EM algorithm.

1. Set $n = 1$, make an initial guess for $\hat{\mathsf{x}}(0)$.
2. Compute $U_{\mathsf{y}}(x, \hat{\mathsf{x}}^{(n-1)})$.
3. Compute $\hat{\mathsf{x}}^{(n)} = \arg\max_a U_{\mathsf{y}}(a, \hat{\mathsf{x}}^{(n-1)})$.
4. If not happy with $\hat{\mathsf{x}}^{(n)}$, increment n and go back to step 2.