Lecture 5 — Sept. 17th, 2021

Sebastien Roch, UW-Madison *Scribe: Liu Yang*

1 Overview

In the last lecture we talked about the bias-variance decomposition in the mean squared loss, and introduce the Cramer-Rao bound.

In this lecture we will prove the Cramer-Rao bound.

2 Unbiased Estimator

Definition 1. *The bias of $\hat{\theta}$ (with respect to distribution P) is*

$$\text{bias}_P(\hat{\theta}) = \mathbb{E}_P(\hat{\theta}) - \theta(P)$$

we say that $\hat{\theta}$ is unbiased if $\mathbb{E}_P(\hat{\theta}) = \theta(P)$, $\forall P \in \mathcal{P}$.

Example: Let the true parameter be θ^*. Defined $\hat{\gamma}^{(n)}(X_1, \cdots, X_n) = g(\theta^*)$, where $g(\theta^*) \neq \theta^*$. Then $g(\theta^*)$ is a biased estimator since

$$\mathbb{E}_{P(\cdot,\theta^*)}[\hat{\gamma}^{(n)}] = g(\theta^*).$$

We will use the concept "unbiased estimator" in the Cramer-Rao bound.

3 Cramer-Rao Bound (Special Case)

In the following, we will give the statement of Cramer-Rao bound for θ dimension $p = 1$ (θ is a scalar).

First, we define the required notation:

1. $\mathcal{X}$ is a finite sample space.
2. The parameter space $\Theta \subseteq \mathbb{R}$ is open.
3. $\mathcal{P} = \{P(\cdot, \theta), \theta \in \Theta\}$ where $P(x, \theta)$ is the probability of observing sample x.
4. $\frac{\partial}{\partial \theta} P(x, \theta)$ exists $\forall x, \theta$ (i.e. $P(x; \theta)$ is continuously differentiable for all x w.r.t. θ).
5. $x_1, \cdots, x_n$ iid $\sim P(\cdot, \theta)$.

6. $P(x, \theta) > 0 \quad \forall x, \theta$

Theorem 2. ***Cramer-Rao Bound.*** *If $\hat{\gamma}^{(n)}(\boldsymbol{x})$ is an unbiased estimator of $g(\theta)$ where g is continuous and differentiable. Then*

$$\underbrace{\mathrm{Var}(\hat{\gamma}^{(n)}(\boldsymbol{x}))}_{\mathrm{MSE}(\hat{\gamma}^{(n)}(\boldsymbol{x}))} \geq \frac{[g'(\theta)]^2}{n \ \underbrace{\mathbb{E}\left[\left(\frac{\partial}{\partial \theta} \log P(x_1, \theta)\right)^2\right]}_{\textit{Fisher Information matrix } I(\theta)}} \tag{1}$$

Remark 3. *The $\mathrm{Var}(\hat{\gamma}^{(n)}(\boldsymbol{x}))$ is same as $\mathrm{MSE}(\hat{\gamma}^{(n)}(\boldsymbol{x}))$. This is because MSE of an estimator can be decomposed into mean and variance:*

$$\mathrm{MSE}(\hat{\gamma}^{(n)}) = \mathrm{bias}(\hat{\gamma}^{(n)})^2 + \mathrm{Var}(\hat{\gamma}^{(n)})$$

As $\hat{\gamma}^{(n)}$ is an unbiased estimator, we know $\mathrm{bias}(\hat{\gamma}^{(n)}) = 0$, thus

$$\mathrm{MSE}(\hat{\gamma}^{(n)}) = \mathrm{Var}(\hat{\gamma}^{(n)}).$$

Remark 4. *The point of finding a lower bound for $\mathrm{Var}(\hat{\gamma}^{(n)}(\boldsymbol{x}))$ is: if we successfully prove the upper bound is same as the lower bound, we can stop looking for a better estimator any more.*

Proof. Let $\boldsymbol{x} = (x_1, \cdots, x_n)$ with $x_i \in \mathcal{X}$, $P^{(n)}(\boldsymbol{x}, \theta) = \prod_{i=1}^n P(x_i, \theta)$. Recall the Cauchy-Schwarz inequality:

$$\begin{aligned}
[\mathrm{Cov}(X, Y)]^2 &= [\mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right]]^2 \\
&= [\langle X - \mathbb{E}[X], Y - \mathbb{E}[Y] \rangle]^2 \\
&\leq \langle X - \mathbb{E}[X], X - \mathbb{E}[X] \rangle \langle Y - \mathbb{E}[Y], Y - \mathbb{E}[Y] \rangle \quad \textit{(apply the Cauchy-Schwarz inequality)} \\
&= \mathbb{E}\left((X - \mathbb{E}[X])^2\right) \mathbb{E}\left((Y - \mathbb{E}[Y])^2\right) \\
&= \mathrm{Var}(X)\,\mathrm{Var}(Y)
\end{aligned}$$

Then for any $\Psi(\boldsymbol{x}, \theta)$,

$$[\mathrm{Cov}(\hat{\gamma}^{(n)}(\boldsymbol{x}), \Psi(\boldsymbol{x}, \theta))]^2 \leq \mathrm{Var}(\hat{\gamma}^{(n)}(\boldsymbol{x}))\,\mathrm{Var}(\Psi(\boldsymbol{x}, \theta))$$

which implies

$$\mathrm{Var}(\hat{\gamma}^{(n)}(\boldsymbol{x})) \geq \frac{[\mathrm{Cov}(\hat{\gamma}^{(n)}(\boldsymbol{x}), \Psi(\boldsymbol{x}, \theta))]^2}{\mathrm{Var}(\Psi(\boldsymbol{x}, \theta))} \tag{2}$$

Choose $\Psi(\boldsymbol{x}, \theta) = \frac{\partial}{\partial \theta} \log P^{(n)}(\boldsymbol{x}, \theta) = \frac{\frac{\partial}{\partial \theta} P^{(n)}(\boldsymbol{x}, \theta)}{P^{(n)}(\boldsymbol{x}, \theta)}$.

Then

$$\begin{aligned}
\mathbb{E}[\Psi(\boldsymbol{x}, \theta)] &= \sum_{\boldsymbol{x} \in \mathcal{X}^n} \cancel{P^{(n)}(\boldsymbol{x}, \theta)} \frac{\frac{\partial}{\partial \theta} P^{(n)}(\boldsymbol{x}, \theta)}{\cancel{P^{(n)}(\boldsymbol{x}, \theta)}} \\
&= \frac{\partial}{\partial \theta} \underbrace{\left(\sum_{\boldsymbol{x} \in \mathcal{X}^n} P^{(n)}(\boldsymbol{x}, \theta) \right)}_{1} \\
&= 0
\end{aligned}$$

And

$$\mathrm{Var}(\Psi(\boldsymbol{x},\theta)) = I_n(\theta) = nI(\theta)$$

is the denominator part in Eq. 1. Recall the definition of covariance matrix:

$$\begin{aligned}
\mathrm{Cov}(X,Y) &= \mathbb{E}\left[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])\right] \\
&= \mathbb{E}\left[XY - Y\mathbb{E}[X] - X\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y]\right] \\
&= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]
\end{aligned}$$

So

$$\begin{aligned}
\mathrm{Cov}(\hat{\gamma}^{(n)}(\boldsymbol{x}), \Psi(\boldsymbol{x},\theta)) &= \mathrm{Cov}\left(\hat{\gamma}^{(n)}(\boldsymbol{x}), \frac{\partial}{\partial\theta}\log P^{(n)}(\boldsymbol{x},\theta)\right) \\
&= \mathbb{E}\left[\hat{\gamma}^{(n)}(\boldsymbol{x})\cdot\frac{\partial}{\partial\theta}\log P^{(n)}(\boldsymbol{x},\theta)\right] - \mathbb{E}[\hat{\gamma}^{(n)}(\boldsymbol{x})]\cdot\underbrace{\mathbb{E}\left[\frac{\partial}{\partial\theta}\log P^{(n)}(\boldsymbol{x},\theta)\right]}_{0} \\
&= \sum_{\boldsymbol{x}\in\mathcal{X}^n} \cancel{P^{(n)}(\boldsymbol{x},\theta)}\cdot\hat{\gamma}^{(n)}(\boldsymbol{x})\cdot\frac{\frac{\partial}{\partial\theta}P^{(n)}(\boldsymbol{x},\theta)}{\cancel{P^{(n)}(\boldsymbol{x},\theta)}} \\
&= \frac{\partial}{\partial\theta}\underbrace{\left(\sum_{\boldsymbol{x}\in\mathcal{X}^n}\hat{\gamma}^{(n)}(\boldsymbol{x})\cdot P^{(n)}(\boldsymbol{x},\theta)\right)}_{\mathbb{E}[\hat{\gamma}^{(n)}(\boldsymbol{x})]} \\
&= \frac{\partial}{\partial\theta}\left(g(\theta)\right) \quad (\textit{since } \hat{\gamma}^{(n)}(\boldsymbol{x}) \textit{ is an unbiased estimator of } g(\theta)) \\
&= g'(\theta)
\end{aligned}$$

We can complete the proof by plugging in the value of $\mathrm{Cov}(\hat{\gamma}^{(n)}(\boldsymbol{x}), \Psi(\boldsymbol{x},\theta))$ to Eq. 2. □

Remark 5. *If the sample space is continuous, then we cannot take the derivative $\frac{\partial}{\partial\theta}$ outside the sum.*

4 Example of Cramer-Rao Bound on Bernoulli Estimator

Let $x = 1$ with probability θ, and $x = 0$ with probability $1-\theta$. We want to calculate the lower bound of any estimator $\hat{\gamma}^{(n)}(\boldsymbol{x})$. Set

- the sample space to be $\mathcal{X} = \{0,1\}$,
- and the parameter space $\Theta = \{0,1\}$,
- $P(x,\theta) = \theta^x(1-\theta)^{1-x}$,
- $g(\theta) = \theta$.

Then

$$
\begin{aligned}
\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log P(x_1,\theta)\right)^2\right] &= \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\left[x_1\log\theta+(1-x_1)\log(1-\theta)\right]\right)^2\right] \\
&= \mathbb{E}\left[\left(\frac{x_1}{\theta}-\frac{1-x_1}{1-\theta}\right)^2\right] \\
&= \theta\cdot\frac{1}{\theta^2}+(1-\theta)\cdot\frac{1}{(1-\theta)^2} \\
&= \frac{1}{\theta}+\frac{1}{1-\theta}=\frac{1}{\theta(1-\theta)}
\end{aligned}
$$

and $g'(\theta)=1$, thus

$$
\mathrm{Var}(\hat{\gamma}^{(n)}(\boldsymbol{x})) \geq \frac{\theta(1-\theta)}{n}
$$

If we estimate by $\hat{\theta}(\boldsymbol{x})=\frac{\sum_{i=1}^n x_i}{n}$, then the variance is exactly the $\frac{\theta(1-\theta)}{n}$. Also this is an unbiased estimator. So by Cramer-Rao bound, we confirm this is the best estimator we can get.

Other choice of $g(\theta)$. If we take $g(\theta)=\frac{1}{\theta}$, then we can show that there is no unbiased estimator, and as $\theta\to 0$, $\mathbb{E}[\hat{\gamma}^{(n)}(\boldsymbol{x})]\approx\hat{\gamma}^{(n)}(\mathbf{0})$.