ECE598: Information-theoretic methods in high-dimensional statistics	Spring 2016
Lecture 7: Information bound	
Lecturer: Yihong Wu	*Scribe: Shiyu Liang, Feb 16, 2016 [Ed. Mar 9]*

Recall the Chi-squared divergence and Hammersley-Chapman-Robbins (HCR) bound from last class. Suppose that P, Q are two probability distribution defined on $\mathbb{R}$, that $X \in \mathcal{X}$ is random variable. The Chi-squared divergence is

$$\chi^2(P\|Q) = \sup_{g:\mathcal{X}\to\mathbb{R}} 2\mathbb{E}_P[g(X)] - \mathbb{E}_Q[g^2(X)] - 1.$$

Furthermore, choosing affine function g yields

$$\chi^2(P\|Q) \geq \frac{(\mathbb{E}_P[X] - \mathbb{E}_Q[X])^2}{\mathrm{var}_Q[X]}$$

which gives the HCR bound.

7.1 HCR Lower Bound

We are now continuing on the HCR lower bound from the last class. We here illustrate an example of HCR lower bound on estimation.

Example 7.1 (Estimation)**.** Let $\theta \in \mathbb{R}$ be an unknown, deterministic parameter, and let $X \in \mathbb{R}$ be a random variable, interpreted as a measure of θ or data. Suppose $\hat{\theta}$ is an unbiased estimate of θ based on X. The relationships can be shown as

$$\theta \to X \to \hat{\theta}.$$

The estimation loss $\ell(\theta, \hat{\theta})$ is defined as $l(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$. Let $P = P_{\theta'}$, $Q = P_\theta$, and then the risk is lower bounded by

$$R_\theta(\hat{\theta}) \geq \mathrm{var}_\theta(\hat{\theta}) \geq \frac{(\mathbb{E}_\theta\hat{\theta} - \mathbb{E}_{\theta'}\hat{\theta})^2}{\chi^2(P_{\theta'}\|P_\theta)}.$$

Suppose $\hat{\theta}$ is an unbiased estimate of θ, then

$$R_\theta(\hat{\theta}) \geq \sup_{\theta\neq\theta} \frac{(\theta - \theta')^2}{\chi^2(P_{\theta'}\|P_\theta)} \geq \lim_{\theta'\to\theta} \frac{(\theta' - \theta)^2}{\chi^2(P_{\theta'}\|P_\theta)}.$$

7.2 Fisher Information

The Fisher information is a way of measuring the amount of information that an observable random variable X carries about an unknown, deterministic parameter θ upon which the probability of the observatoin X depends. Assume the probability density function of random variable X conditional on the value of θ is p_θ. The Fisher information is defined as

Definition 7.1 (Fisher Information)**.** The Fisher information of the parameteric family of densitities $\{p_\theta : \theta \in \Theta\}$ (with respect to μ) at θ is

$$I(\theta) = \mathbb{E}\left[\left(\frac{\partial \log p_\theta}{\partial \theta}\right)^2\right] = \int \left(\frac{\partial p_\theta}{\partial \theta}\right)^2 \cdot \frac{1}{p_\theta}. \tag{7.1}$$

Theorem 7.1 (Fisher Information)**.** *If p_θ is twice differentiable with respect to θ, the Fisher information can be written as*

$$I(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial \theta^2} \log p_\theta\right]$$

Proof. Since

$$\frac{\partial^2}{\partial \theta^2} \log p_\theta = \frac{\frac{\partial^2}{\partial \theta^2} p_\theta}{p_\theta} - \left(\frac{\frac{\partial}{\partial \theta} p_\theta}{p_\theta}\right)^2 = \frac{\frac{\partial^2}{\partial \theta^2} p_\theta}{p_\theta} - \left(\frac{\partial}{\partial \theta} \log p_\theta\right)^2$$

and

$$\mathbb{E}\left[\frac{\partial^2 p_\theta}{\partial \theta^2} \cdot \frac{1}{p_\theta}\right] = \frac{\partial^2}{\partial \theta^2} \int p_\theta \mu(dx) = \frac{\partial^2}{\partial \theta^2} 1 = 0.$$

Thus, we have

$$I(\theta) = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial \theta} \log p_\theta\right)^2\right] = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial \theta^2} \log p_\theta\right].$$

□

Theorem 7.2 (Fisher Information: mutiple sample)**.** *Suppose random sample $X_1, \ldots, X_n$ independently and identically drawn from a distribution p_θ. The Fisher information $I_n(\theta)$ provided by random samples $X_1, \ldots, X_n$ is*

$$I_n(\theta) = nI(\theta),$$

where $I(\theta)$ is Fisher information provided by a single sample X_1.

Proof. We first denote the joint pdf of $X_1, \ldots, X_n$ as

$$p_\theta(x_1, \ldots, x_n) = \prod_{i=1}^n p_\theta(x_i).$$

Then the Fisher information $I_n(\theta)$ provided by $X_1, \ldots, X_n$ is

$$I_n(\theta) = \mathbb{E}_\theta\left[\left(\frac{\partial l_\theta(X_1, \ldots, X_n)}{\partial \theta}\right)^2\right] = \int \cdots \int \left(\frac{\partial l_\theta(x_1, \ldots, x_n)}{\partial \theta}\right)^2 p_\theta(x_1, \ldots, x_n) dx_1 dx_2 \ldots dx_n,$$

which is an n-dimensional integral. Thus, by Theorem 7.1, the Fisher information provided by $X_1, \ldots, X_n$ can be calculated as

$$I_n(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2 \log p_\theta(X_1, \ldots, X_n)}{\partial \theta^2}\right] = -\mathbb{E}_\theta\left[\sum_{i=1}^n \frac{\partial^2 \log p_\theta(X_i)}{\partial \theta^2}\right] = -\sum_{i=1}^n \mathbb{E}_\theta\left[\frac{\partial^2 \log p_\theta(X_i)}{\partial \theta^2}\right] = nI(\theta).$$

□

7.3 Variantions of HCR/CR Lower Bound

This section contains the following three versions of HCP/CR lower bound:

- Multiple Samples Version
- Multivariate Version
- Functional Version

7.3.1 Multiple Samples Version

Suppose θ is some unknown, deterministic parameter and $X_1, \ldots, X_n$ are n random variables independently and identically coming from the distribution P_θ. The estimate $\hat{\theta}$ comes from $X_1, \ldots, X_n$. The relationships is shown as follows:

$$\theta \to X_1, \ldots, X_n \to \hat{\theta}.$$

Then the risk is lower bound by

$$R_\theta(\hat{\theta}) \geq \mathrm{var}_\theta \hat{\theta} \geq \frac{(\mathbb{E}_\theta \hat{\theta} - \mathbb{E}_{\theta'} \hat{\theta})^2}{\chi^2(P_{\theta'}^{\otimes n} \| P_\theta^{\otimes n})}.$$

For the HCR lower bound,

$$R_\theta(\hat{\theta}) \geq \sup_{\theta \neq \theta'} \frac{(\theta - \theta')^2}{(1 + \chi^2(P_\theta \| P_{\theta'}))^n - 1} \overset{\theta' \to \theta}{\geq} \frac{1}{nI(\theta)}.$$

We next show the counterpart for

$$\chi^2(P \| Q) \geq \frac{(\mathbb{E}_P X - \mathbb{E}_Q X)^2}{\mathrm{var}_Q X}.$$

Suppose P, Q are two distributions defined on $\mathbb{R}^p$, then

$$\chi^2(P \| Q) = \sup_{g : \mathbb{R}^p \to \mathbb{R}} [2\mathbb{E}_P g(X) - \mathbb{E}_Q g^2(X) - 1].$$

Furthter, if $g(X) = \langle a, X \rangle + 1$, then

$$\chi^2(P \| Q) \geq 2\mathbb{E}_P \langle a, X \rangle + 1 - \mathbb{E}_Q(\langle a, X \rangle + 1)^2.$$

If we further assume $\mathbb{E}_Q X = 0$, then we have

$$\chi^2(P \| Q) \geq 2 \langle a, \mathbb{E}_P X \rangle - a^T \mathbb{E}_Q[X X^T] a.$$

Therefore, we finally have

$$\chi^2(P \| Q) \geq (\mathbb{E}_P X - \mathbb{E}_Q X)^T \mathrm{cov}_Q^{-1}(X)(\mathbb{E}_P X - \mathbb{E}_Q X)$$

7.3.2 Multivariate Version

Let the loss function $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|_2^2$ and $\hat{\theta}$ be the unbiased estimate of θ, *i.e.*, $\mathbb{E}_\theta \hat{\theta} = \theta$. Then

$$(\theta' - \theta)^T \mathrm{cov}_\theta^{-1}(\hat{\theta})(\theta' - \theta) \leq \chi^2(P_{\theta'} \| P_\theta) \overset{\theta' \to \theta}{=} (\theta' - \theta)^T I(\theta)(\theta' - \theta) + \|\theta' - \theta\|_2^2,$$

where the equality follows from the Taylor expansion and Fisher information matrix is given as

$$I(\theta) = \int \frac{\nabla P_\theta (\nabla P_\theta)^T}{P_\theta}.$$

If we take $\theta' = \theta + \epsilon u$, $\epsilon \to 0$, then we have

$$u^T \mathrm{cov}_\theta^{-1}(\hat{\theta}) u \leq u^T I(\theta) u,$$

which is equivalent to

$$\mathrm{cov}_\theta(\hat{\theta}) \succeq I^{-1}(\theta),$$

and further indicates

$$R_\theta(\hat{\theta}) = \mathrm{tr}(\mathrm{cov}_\theta(\hat{\theta})) \geq \mathrm{tr}(I^{-1}(\theta)).$$

Then we have

$$\mathbb{E}\|\theta - \hat{\theta}\|_2^2 = \sum_{i=1}^p \mathbb{E}(\hat{\theta}_i - \theta_i)^2 \geq \sum_{i=1}^p \frac{1}{I_i},$$

where $I_i = (I(P_\theta)_{ii})$ since

$$\sum_{i=1}^p \frac{1}{I_i(\theta)} \leq \mathrm{tr}(I^{-1}(\theta)).$$

Note that if we apply the one-dimensional CRLB for each coordinate we would get the rightmost inequality which is weaker. In addition, the Fisher information matrix can be written as

$$I(\theta) = \mathbb{E}_\theta[(\nabla \log P_\theta)(\nabla \log P_\theta)^T] = \mathrm{cov}_\theta(\nabla \log P_\theta) = -\left(\mathbb{E}_\theta\left[\frac{\partial^2 \log P_\theta}{\partial\theta_i \partial\theta_j}\right]\right).$$

7.3.3 Functional Version

Assume that θ is an unknown parameter, that random variable X comes from the distribution P_θ and that $\hat{T}(X)$ is an estimation for $T(\theta)$, where $T : \Theta \to \mathbb{R}$. The relationship is shown as follows:

$$\theta \to X \to \hat{T}.$$

If we further assume $\hat{T}(\theta)$ is an unbiased estimation for $T(\theta)$, then

$$\mathrm{var}_\theta(\hat{T}) \geq \frac{\|\nabla T\|_2^2}{I(\theta)}$$

7.4 Bayesian Cramér-Rao Lower Bound

The class will introduce two methods of proving Bayesian Cramér-Rao lower bound.

- Method 1: $\chi^2 \to$ Bayesian HCR $\to$ Bayesian CR
- Method 2: Classical Method

The notation used in this section is shown as follows:

- $\Theta = \mathbb{R}$
- $\ell(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$.
- π is a "nice" prior on $\mathbb{R}$

The relationship can be described as follows:

$$\pi \to \theta \to X \to \hat{\theta}.$$

Theorem 7.3 (Bayesian Cramér-Rao Lower Bound)**.** *Assuming suitable regularity conditions, then*

$$R^* \geq R^*_\pi = \inf_{\hat{\theta}} \mathbb{E}_\pi(\theta, \hat{\theta})^2 \geq \frac{1}{\mathbb{E}_{\theta\sim\pi} I(\theta) + I(\pi)},$$

where R^*_π is the Bayes risk and $I(\pi) = \int \frac{\pi'^2}{\pi}$.

Let

$$Q : \pi \longrightarrow \theta \overset{P_\theta = Q_{X|\theta}}{\longrightarrow} X \longrightarrow \hat{\theta},$$

$$P : \tilde{\pi} \longrightarrow \theta \overset{\tilde{P}_\theta = P_{X|\theta}}{\longrightarrow} X \longrightarrow \hat{\theta}.$$

Then

$$\begin{aligned}
\chi^2(P_{\theta X} \| Q_{\theta X}) &\geq \chi^2(P_{\theta\hat{\theta}} \| Q_{\theta\hat{\theta}}) \leftarrow \text{data processing inequality} \\
&\geq \chi^2(P_{\theta-\hat{\theta}} \| Q_{\theta-\hat{\theta}}) \leftarrow \text{data processing inequality} \\
&\geq \frac{(\mathbb{E}(\theta - \hat{\theta}) - \mathbb{E}_Q(\theta - \hat{\theta}))^2}{\mathrm{var}_\theta(\hat{\theta} - \theta)} \\
&= \frac{\delta^2}{\mathrm{var}_\theta(\theta - \hat{\theta})}
\end{aligned}$$

Further, if we assume

$$Q_\theta = \pi, Q_{X|\theta} = P_\theta, P_\theta = T_\delta \pi, P_{X|\theta} = P_{\theta-\delta},$$

then $P_X = Q_X$ which further indicates $P_{\hat{\theta}} = Q_{\hat{\theta}}$ and the mean of $\hat{\theta}$ under distribution of P equals to the mean under the distribution under Q. For the Bayesian HCR lower bound,

$$R^*_\pi \geq \sup_{\delta\neq 0} \frac{\delta^2}{\chi^2(P_{X\theta} \| Q_{X\theta})} \geq \lim_{\delta\to 0} \frac{\delta^2}{\chi^2(P_{X\theta} \| Q_{X\theta})} = \frac{1}{I(\pi) + \mathbb{E}_{\theta\sim\pi}[I(\theta)]}. \tag{7.2}$$

We give a short proof of (7.2) here.

Proof.

$$
\begin{aligned}
\chi^2(P_{X\theta}\|Q_{X\theta}) &= \int \frac{(P_{X\theta}-Q_{X\theta})^2}{Q_{X\theta}} = \int \frac{[P_\theta(P_{X|\theta}-Q_{X|\theta})+(P_\theta-Q_\theta)Q_{X|\theta}]^2}{Q_{X\theta}} \\
&= \int \frac{P_\theta^2}{Q_\theta}\int \frac{(P_{X|\theta}-Q_{X|\theta})^2}{Q_{X|\theta}} + \int \frac{(P_\theta-Q_\theta)^2}{Q_\theta^2} + 2\int \frac{P_\theta(P_\theta-Q_\theta)}{Q_\theta}\int(P_{X|\theta}-Q_{X|\theta}) \\
&= \chi^2(P_\theta\|Q_\theta) + \mathbb{E}\left[\chi^2(P_{X|\theta}\|Q_{X|\theta})\cdot\left(\frac{P_\theta}{Q_\theta}\right)^2\right]
\end{aligned}
$$

Then applying

- $\chi^2(P_\theta\|Q_\theta) = \chi^2(T_{\delta\pi}\|\pi) = \delta^2[I(\pi)+o(1)]$ by Taylor expansion,
- $\chi^2(P_{X|\theta}\|Q_{X|\theta}) = [I(\theta)+o(1)]\delta^2$ by Taylor expansion,

we obtain (7.2). □

7.5 Information Bound

In this section, we introduce the local version of the minimax lower bound. The local minimax risks is defined in a quadratic form: $\inf_{\hat{\theta}}\sup_{|\theta-\theta_0|\le\epsilon}\mathbb{E}(\hat{\theta}-\theta)^2$. Further, we have

$$
\begin{aligned}
\inf_{\hat{\theta}}\sup_{|\theta-\theta_0|\le\epsilon}\mathbb{E}(\hat{\theta}-\theta)^2 &\ge \frac{1}{I(\theta)+n\mathbb{E}_{\theta\sim\pi}[I(\theta)]} \\
&= \frac{1+o(1)}{n\mathbb{E}_{\theta\sim\pi}[I(\theta)]}
\end{aligned}
$$

If $\theta\mapsto I(\theta)$ is continuous, then

$$
\mathbb{E}_{\theta\sim\pi}[I(\theta)] = I(\theta_0)+o(1) = \frac{1+o(1)}{nI(\theta)}.
$$

Assume the random variable Z coming from the distribution π, $Z\sim\pi$. Let $I(Z)\triangleq I(\pi)$. For constant $\alpha,\beta\neq 0$, then $I(Z+\alpha)=I(Z)$ and $I(\beta Z)=\frac{I(Z)}{\beta^2}$. If the π has the distribution of form $\cos^2\frac{\pi x}{2}$, then $\min_{\pi:[-1,1]}I(\pi)=\pi^2$. If the distribution π has the form of $\cos^2\frac{\pi(x-\theta_0)}{2\epsilon}$, then $I(\theta)=\frac{\pi^2}{\epsilon}$. Then we have

$$
\inf_{\hat{\theta}}\sup_{|\theta_0-\theta|\le\epsilon}\mathbb{E}(\hat{\theta}-\theta)^2 \ge R_\pi^* \ge \frac{1}{n\mathbb{E}_{\theta\sim\pi}[I(\theta)]+I(\pi)}.
$$

Now if we pick $\epsilon=n^{-1/4}$, we have

$$
R^* \ge \inf_{\hat{\theta}}\sup_{|\theta-\theta_0|\le n^{-1/4}}\mathbb{E}_\theta(\theta-\hat{\theta})^2 \ge \frac{1}{nI(\theta)+o(\sqrt{n})} \overset{Optimize}{\Longrightarrow} R^* \ge \frac{1+o(1)}{n\inf_{\theta_0\in\Theta}I(\theta_0)}.
$$