10-704: Information Processing and Learning **Spring 2012**

Lecture 23: Cramér-Rao and Uninformative Priors

Lecturer: Aarti Singh *Scribes: Soumya Batra*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

23.1 Cramer Rao Lower Bound

This is a technique for lower bounding the performance of unbiased estimators. Let $p(x;\theta)$ be a probability density function with continuous parameter $\theta \in \Theta$. Let $X_1, \ldots, X_n$ be n i.i.d samples from this distribution, i.e., $X_i \sim p(x;\theta)$. Let $\hat{\theta}(X_1, \ldots, X_n)$ be an unbiased estimator of θ, so that $\mathbb{E}\hat{\theta} = \theta$ for all $\theta \in \Theta$.

Now, if $p(x;\theta)$ satisfies the following two conditions:

1. $$\frac{\partial}{\partial\theta}\left[\int\ldots\int \hat{\theta}(x_1,\ldots,x_n)\prod_{i=1}^{n} p(x_i;\theta)\right] = \int\ldots\int \hat{\theta}(x_1,\ldots,x_n)\frac{\partial\prod_{i=1}^{n} p(x_i;\theta)}{\partial\theta}dx_1\ldots dx_n \qquad (23.1)$$

   This is a fairly mild continuity condition, allowing us to push the derivative through the integrals.

2. For each θ, the variance of $\hat{\theta}(X_1, \ldots, X_n)$ is finite.

then the variance of the unbiased estimator is bounded as:

$$\text{var}(\hat{\theta}) \geq \frac{1}{n\mathbb{E}_X\left[\left(\frac{\partial \log p(x;\theta)}{\partial\theta}\right)^2\right]} = \frac{1}{-n\mathbb{E}_X\left[\frac{\partial^2 \log p(x;\theta)}{\partial\theta^2}\right]} = \frac{1}{I(\theta)},$$

where $I(\theta)$ is the Fisher Information. The Fisher information characterizes the curvature of the log likelihood function. CR lower bound states that larger the curvature, the smaller is the variance since the likelihood changes sharply around the true parameter.

Moreover, this bound is achieved for all θ if the following condition is met:

$$\forall\theta, \qquad \frac{\partial}{\partial\theta}\log(p(x;\theta)) = I(\theta)(\hat{\theta}(x) - \theta)$$

We can see that this is an important result as now we are able to bound the variance of a specific unbiased estimator of our choice rather than having a general bound over all possible estimators as in the minimax lower bounds.

23.2 Proof

We will prove the Cramer-Rao Lower Bound for $n = 1$. We can prove the bound similarly for a more general case with $n > 1$ (See [1]). Since we are considering unbiased estimators:

$$0 = \mathbb{E}_{p(x;\theta)}[\hat{\theta} - \theta] = \int \Big(\hat{\theta}(x) - \theta\Big)\, p(x;\theta) dx$$

Differentiating both sides w.r.t θ and using Equation 23.1 we get

$$\begin{aligned}
0 = \frac{\partial}{\partial\theta}\left[\int \Big(\hat{\theta}(x) - \theta\Big)\, p(x;\theta)\right] &= \int \frac{\partial}{\partial\theta}\left[\Big(\hat{\theta}(x) - \theta\Big)\, p(x;\theta)\right] dx \\
&= \int \Big(\hat{\theta}(x) - \theta\Big)\frac{\partial p(x;\theta)}{\partial\theta} + p(x;\theta)\underbrace{\frac{\partial}{\partial\theta}\Big(\hat{\theta}(x) - \theta\Big)}_{=-1(\text{since } \hat{\theta} \perp\!\!\!\perp \theta)} dx \\
&= \int \Big(\hat{\theta}(x) - \theta\Big)\frac{\partial p(x;\theta)}{\partial\theta} dx + \int p(x;\theta)(-1) dx \\
&= \int \Big(\hat{\theta}(x) - \theta\Big)\frac{\partial p(x;\theta)}{\partial\theta} dx - \underbrace{\int p(x;\theta) dx}_{=1} \\
&= \int \Big(\hat{\theta}(x) - \theta\Big)\, p(x;\theta)\frac{\partial \log\left(p\left(x;\theta\right)\right)}{\partial\theta} dx - 1 \qquad \left(\text{using identity } \frac{\partial \log f}{\partial g} = \frac{1}{f}\frac{\partial f}{\partial g}\right) \\
\text{or} \qquad 1 &= \int \Big(\hat{\theta}(x) - \theta\Big)\sqrt{p(x;\theta)}\sqrt{p(x;\theta)}\frac{\partial \log\left(p\left(x;\theta\right)\right)}{\partial\theta} dx
\end{aligned}$$

Taking square of both sides,

$$1 = \left[\int \underbrace{\Big(\hat{\theta}(x) - \theta\Big)\sqrt{p(x;\theta)}}_{f}\underbrace{\sqrt{p(x;\theta)}\frac{\partial \log\left(p\left(x;\theta\right)\right)}{\partial\theta}}_{g} dx\right]^2$$

Applying Cauchy-Schwartz inequality ($(\int fg)^2 \le \int f \cdot \int g$) on RHS assuming the 2 functions to be f and g as shown above,

$$1 \le \underbrace{\int \Big(\hat{\theta}(x) - \theta(x)\Big)^2 p(x;\theta) dx}_{var(\hat{\theta})}\underbrace{\int p(x;\theta)\left[\frac{\partial \log\left(p\left(x;\theta\right)\right)}{\partial\theta}\right]^2 dx}_{\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log p(x;\theta)\right)^2\right]}$$

Rearranging, we get the Cramer-Rao Lower bound for a single sample case.

23.2.1 Note: when θ is multi-dimensional

In this case, Fisher's Information is a matrix where $[I(\theta)]_{ij} = -n\mathbb{E}[\frac{\partial^2}{\partial\theta_i \partial\theta_j}\log p(x;\theta)]$. Thus,

$$\begin{aligned}
& cov(\hat{\theta}) \succeq \frac{1}{I(\theta)} \qquad \text{where } A \succeq B \text{ denotes that } A - B \text{ is positive semi-definite.} \\
\implies \quad & var(\hat{\theta}) \ge \frac{1}{tr(I(\theta))}
\end{aligned}$$

23.3 Remarks

We note the following points with respect to Cramer-Rao Lower Bound (CRLB).

1. Both conditions on $p(x;\theta)$ are necessary for the bound to hold. For example, condition 1 does not hold for the uniform distribution $U(0,\theta)$ and hence the CRLB is not valid. In other cases (e.g. if condition 2 does not hold), the CRLB bound may be too loose (sometimes just stating $var(\hat{\theta}) \geq 0$).
2. CRLB holds for a specific estimator $\hat{\theta}$ and does not give a general bound on all estimators.
3. CRLB applies to unbiased estimators alone, though a version that extends it to biased estimators also exists, which we will see soon. Hence, it is useful for parametric problems (where unbiased estimator typically have the same rate of convergence as the minimax optimal estimator) but not usually for non-parametric or high-dimensional ($d \gg n$) problems (where the bias and variance tradeoff plays an important role in determining the rate of convergence).

23.4 Examples

We see a few examples where CRLB is applicable. Assume n samples in each case.

23.4.1 Gaussian: $\mathcal{N}(\mu, \sigma^2)$

- The sample mean, $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_n$, for n samples is an unbiased estimator of the mean. This attains CRLB for Gaussian mean and calculation of the Fisher information shows that $var(\hat{\mu}) \geq \frac{\sigma^2}{n}$ for n samples.
- Sample median, on the other hand, is an unbiased estimator of the mean that does not attain CRLB.

23.4.2 Least Squares in Linear Regression model : $X = A\theta + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2)$

The Least Squares solution $\hat{\theta} = (A^T A)^{-1} A^T X$ is unbiased in low-dimensional settings and attains CRLB.

Calculation of the Fisher information reveals that $I(\theta) = \frac{A^T A}{\sigma^2}$ and hence $cov(\hat{\theta}) \succeq \sigma^2 (A^T A)^{-1}$.

23.4.3 Gaussian with known mean : $\mathcal{N}(\mu, \sigma^2)$

Sample unbiased estimator for variance: $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$.

Its variance is calculated as $var(\hat{\theta}) = \frac{2\sigma^4}{n}$.

We get the CRLB bound as: $var(\hat{\theta}) \geq \frac{2\sigma^4}{n}$.

Hence, we see that this estimator attains CRLB.

Note that if we had considered a biased estimator, it could result in smaller variance and hence smaller mean square error (for unbiased estimators, mean square error is just the variance).

For example, $\hat{\theta}_{\text{biased}} = \frac{1}{n+2}\sum_{i=1}^{n}(X_i - \mu)^2$. This has variance $var(\hat{\theta}_{\text{bias}}) = \frac{2n\sigma^4}{(n+2)^2}$, clearly less than the CRLB bound.

The bias for this estimator $= \frac{2\sigma^2}{n+2}$. $\implies$ Mean Squared Error $= \frac{2\sigma^4}{n+2}$, which is a constant improvement over unbiased estimators.

23.5 Extension of CRLB to biased estimators

CRLB is also extended to work with biased estimators. However, this bound is not widely used.

$$cov(\hat{\theta}) \succeq \left(\left(\mathbb{I} + \frac{db(\theta)}{d\theta}\right) I^{-1}(\theta) \left(\mathbb{I} + \frac{db(\theta)}{d\theta}\right)^T\right)$$

where $\mathbb{I}$ is the identity matrix and $b(\theta)$ is the bias.

23.6 Priors

The Fisher information also plays a role in defining uninformative priors. We discuss this next. We will discuss a few strategies of coming up with priors for a distribution. These are especially important when data is small, resulting in small posterior probabilities.

Consider a parameter $\theta \in \Theta$. We will represent the prior distribution on θ as $\pi(\theta)$.

23.6.1 Uninformative prior

Often priors that don't influence the posterior distribution too much are preferred. Such priors are called uninformative priors. A common uninformative prior is an unbiased flat prior which assigns constant probability to all $\theta \in \Theta$, i.e. the uniform distribution.

It suffers from 2 main problems:

- It is not invariant, i.e., posterior probabilities using the flat prior and some transformation of it may lead to different inferences.

  Eg. Consider $X \sim Bin(n, \theta), \pi(\theta) = 1, 0 \leq \theta \leq 1$.
  Now, consider a log transformation over parameter θ: $\phi = h(\theta) = \log \frac{\theta}{1-\theta}$. Notice that the support for ϕ has become infinite as well as ϕ has become informative now (no longer a flat prior).

- It may become biased in high dimensions.

  Eg. As the number of dimensions $d \to \infty$, most of the mass of a uniform distribution on the d-dimensional hypercube starts to lie at ∞. In such a setting, a Gaussian distribution which is uniform on any d-dimensional sphere might be more appropriate.

23.6.2 Jeffrey's prior

Jeffrey's prior improves upon the flat prior by being invariant in nature. To understand invariance, lets consider the posterior on which inferences are based. For θ, if $\pi(\theta)$ is the prior, then by Bayes rule the

posterior is $p(\theta|x) \propto \pi(\theta)p(x;\theta)$ where $p(x;\theta)$ is the likelihood of observing data x under θ. If we now consider a transformation $\phi = h(\theta)$, then using rules for transformation of random variables $p(\phi|x) = p(\theta|x)\left|\frac{d\theta}{d\phi}\right| \propto p(\theta)\left|\frac{d\theta}{d\phi}\right| p(x;\theta)$. By Bayes rule, $p(\phi|x) \propto \pi(\phi)p(x;\phi) = \pi(\phi)p(x;\theta)$. Hence, for invariance we must have

$$\pi(\phi) = \pi(\theta)\left|\frac{d\theta}{d\phi}\right|.$$

The Jeffrey's prior is a non-informative prior distribution on parameter space that is proportional to the determinant of Fisher information,i.e.,

$$\pi_J(\theta) \propto \sqrt{|I(\theta)|}$$

where $|\cdot|$ denotes the determinant.

For a single dimension case, this is simply

$$\pi_J(\theta) \propto \sqrt{I(\theta)}$$

Jeffrey's prior is usually used for single dimension cases as in higher dimensions, it starts to bias the solution by a large factor.

Consider a transformation over parameter $\theta : \phi = h(\theta)$ in a one-dimension case. Then, we have

$$\sqrt{I(\phi)} = \sqrt{I(\theta)}\left|\frac{d\theta}{d\phi}\right|$$

This ensures the invariant property of the Jeffrey's prior.

Proof: Beginning with RHS,

$$\begin{aligned}\sqrt{I(\theta)}\left|\frac{d\theta}{d\phi}\right| &= \sqrt{I(\theta)\left(\frac{d\theta}{d\phi}\right)^2} = \sqrt{\mathbb{E}\left[\left(\frac{d\log p(x;\theta)}{d\theta}\right)^2\right]\left(\frac{d\theta}{d\phi}\right)^2} \\ &= \sqrt{\mathbb{E}\left[\left(\frac{d\log p(x;\theta)}{\cancel{d\theta}}\frac{\cancel{d\theta}}{d\phi}\right)^2\right]} = \sqrt{\mathbb{E}\left[\left(\frac{d\log p(x;\theta)}{d\phi}\right)^2\right]} = \sqrt{I(\phi)}\end{aligned}$$

Hence, invariant property is maintained. For higher dimensions, replace $I(\theta)$ with $|I(\theta)|$ in the above equations.

23.6.2.1 Examples

Example 1: Jeffrey's prior for mean of Gaussian distribution: $\mathcal{N}(\mu, \sigma^2)$

Consider posterior : $p(x;\mu)$

$$\begin{aligned} & p(x;\mu) \propto e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \\ \implies \quad & \frac{\partial^2}{\partial\mu^2}\log(p(x;\mu)) \propto -\frac{1}{\sigma^2} \\ \implies \quad & I(\mu) \propto \frac{1}{\sigma^2}\end{aligned}$$

which is independent of μ. Thus, for a fixed σ^2,

$$\pi_J(\mu) \propto 1 \qquad \text{Flat prior}$$

Similarly, we can calculate the Jeffrey's prior for the variance and it turns out that

$$\pi_J(\sigma) \propto \frac{1}{\sigma}$$

which is not a flat prior. In fact, it is flat for the transformed variable $\phi = \log \sigma$. This makes intuitive sense since if we don't know the scale (variance) of a parameter, then a scale of 1:10 should be as likely as a scale of 10:100.

Example 2: Binomial Distribution : $\mathbf{X \sim B(n, \theta), 0 \leq \theta \leq 1}$

Liikelihood $p(x;\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$

$$\begin{aligned}
\pi_J(\theta) \propto \sqrt{I(\theta)} &= \sqrt{-\mathbb{E}\left[\frac{d^2}{d\theta^2}\log p(x;\theta)\right]} = \sqrt{-\mathbb{E}\left[\frac{d}{d\theta}\left(\frac{x}{\theta} - \frac{n-x}{1-\theta}\right)\right]} \\
&= \sqrt{-\mathbb{E}\left[-\frac{x}{\theta^2} - \frac{n-x}{(1-\theta)^2}\right]} = \sqrt{\frac{n\theta}{\theta^2} + \frac{n-n\theta}{(1-\theta)^2}} = \sqrt{\frac{n}{\theta(1-\theta)}} \\
\implies \quad \pi_J(\theta) &\propto \frac{1}{\sqrt{\theta(1-\theta)}}
\end{aligned}$$

which is again inversely proportional to the standard deviation, implying the prior is uniform on log standard deviation.

23.6.3 Reference Priors

In multi-dimensional settings, reference priors are more useful. Motivation for Reference priors lies in formalizing the notion that the prior should not influence the posterior once enough data has been observed.

It is defined as the maximum over some measure of distance or divergence between posterior and prior distributions. Since data is not observed when defining a prior, we can instead use the Expectation. Let us calculate Reference prior for KL-divergence measure over data taking expectation.

$$\begin{aligned}
\pi_R(\theta) &= \arg\max_{\pi(\theta)} \mathbb{E}_X\left[\mathrm{KL}\left(p(\theta|X)\|p(\theta)\right)\right] \\
&= \arg\max_{\pi(\theta)} \int p(x)\int p(\theta|x)\log\frac{p(\theta|x)}{\pi(\theta)}d\theta dx \\
&= \arg\max_{\pi(\theta)} \int p(x,\theta)\log\frac{p(\theta,x)}{p(x)\pi(\theta)}d\theta dx \\
&= \arg\max_{\pi(\theta)} I(\theta;X) \qquad \text{Mutual Information between } \theta \text{ and } X
\end{aligned}$$

This is precisely the capacity of the data-generating channel i.e. the channel with θ as input and X as output.

Since Mutual Information is invariant under reparameterization, we can see that Reference prior too is invariant.

For one-dimensions, reference prior is same as Jeffrey's prior, but it is often different and less biased in multi-dimensions.

23.6.3.1 Connection to Redundancy Capacity Theorem

Recall that we had discussed the redundancy capacity theorem which states that the worst case Bayesian redundancy is same as minimax redundancy

$$\sup_{\pi(\theta)} \inf_{q} \int \pi(\theta) \mathrm{KL}(p_\theta \| q) d\theta = \inf_{q} \sup_{\pi(\theta)} \int \pi(\theta) \mathrm{KL}(p_\theta \| q) d\theta$$

and is equal to $\sup_{\pi(\theta)} I(\theta; X)$ for the optimal $q = q^*$ where q^* is a mixture over p_θ with weights $\pi(\theta)$.

This implies that the worst case prior for Bayesian redundancy is the reference prior, which is the Capacity achieving prior.

References

[1] Adam Merberg and Steven J. Miller, "The Cramer-Rao Inequality," *Course Notes for Math 162: Mathematical Statistics*, 2008.

[2] Michael I. Jordan, "Jeffreys Priors and Reference Priors," *Lecture 7: Stat260: Bayesian Modeling and Inference*, 2010.

[3] "http://en.wikipedia.org/wiki/Jeffreys_prior"