10-704: Information Processing and Learning **Fall 2016**

Lecture 24: Dec 7

Lecturer: Aarti Singh

Note: *These notes are based on scribed notes from Spring15 offering of this course. LaTeX template courtesy of UC Berkeley EECS dept.*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

24.1 Error exponents in Hypothesis Testing

In this lecture, we continue our discussion of using Sanov's theorem to get error exponents in hypothesis testing. Recall that the optimal test is the log likelihood ratio test which can be stated as follows in terms of information theoretic quantities:

$$\begin{aligned}\text{Log likelihood ratio} &= \log \frac{P_0(x^n)}{P_1(x^n)} \\ &= n[D(P_{x^n}||P_1) - D(P_{x^n}||P_0)]\end{aligned}$$

Thus, the decision region corresponding to the likelihood ratio test can be written as:

$$A(T) = \left\{ x^n : D(P_{x^n}||P_1) - D(P_{x^n}||P_0) > \frac{1}{n}\log T \right\}$$

i.e. it is the region of the probability simplex bounded by the set of types for which the difference of the KL divergence to the distributions under the two hypotheses is a constant, i.e. the boundary is parallel to the perpendicular bisector of the line connecting P_0 and P_1. See Figure 24.1.

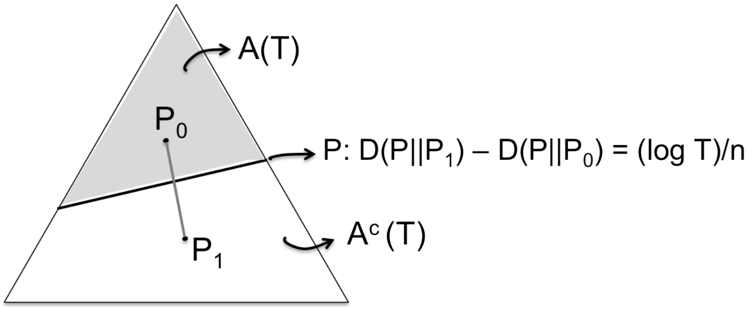

Figure 24.1: The decision region corresponding to a likelihood ratio test is demarcated by boundary that is parallel to the perpendicular bisector of the line joining the distributions under the two hypotheses P_0 and P_1.

To study how the probability of error decays as a function of n in hypothesis testing, we will use large deviation theory (what is the probability that an empirical observation deviates from the true value).

24.1.1 Error-exponents

Using Sanov's theorem, we get that asymptotically the probability of false alarm (type I error)

$$\alpha = P_0(A^c) \approx 2^{-nD(P_0^*||P_0)}$$

where $P_0^* = \arg\min_{P \in A^c} D(P||P_0)$ and

$$\beta = P_1(A) \approx 2^{-nD(P_1^*||P_1)}$$

where $P_1^* = \arg\min_{P \in A} D(P||P_1)$.

In next class, we will evaluate the form of P_1^* (and P_0^*).

Let us evaluate the form of P_1^* (and P_0^*). Notice from Figure 24.1 that since the decision regions are delineated by a line parallel to the perpendicular bisector, P_0^* (the projection of P_0 onto A^c) is same as P_1^* (the projection of P_1 onto A). So we will derive the form of one of them, say P_1^* (you can check following the same arguments that the form of P_0^* is indeed the same).

To evaluate P_1^*, consider the following constrained optimization:

$$\min_P D(P||P_1) \quad s.t. \quad P \in A \equiv D(P||P_1) - D(P||P_0) > \frac{1}{n}\log T$$

Forming the Lagrangian where $\lambda > 0$ and ν are Lagrange multipliers (notice that since we require $\lambda > 0$, we consider the constraint written with a $<$ instead of $>$):

$$\begin{aligned} L(P, \lambda, \nu) &= D(P||P_1) + \lambda(D(P||P_0) - D(P||P_1)) + \nu \sum P \\ &= \sum_{x^n} P(x^n) \log \frac{P(x^n)}{P_1(x^n)} + \lambda \sum_{x^n} P(x^n) \log \frac{P_1(x^n)}{P_0(x^n)} + \nu \sum_{x^n} P(x^n) \end{aligned}$$

Taking the derivative with respect to $P(x^n)$:

$$\log \frac{P(x^n)}{P_1(x^n)} + 1 + \lambda \log \frac{P_1(x^n)}{P_0(x^n)} + \nu \Bigg|_{P=P_1^*} = 0$$

and setting it equal to 0 yields P_1^*:

$$P_1^*(x^n) = e^{-\nu-1} P_0^\lambda(x^n) P_1^{1-\lambda}(x^n) = \frac{P_0^\lambda(x^n) P_1^{1-\lambda}(x^n)}{\sum_{a^n \in \mathcal{X}^n} P_0^\lambda(a^n) P_1^{1-\lambda}(a^n)}$$

where in the last step we substituted for ν by solving for the constraint $\sum_{x^n} P_1^*(x^n) = 1$. In the last expression λ should be chosen to satisfy the constraint $D(P_1^*||P_1) - D(P_1^*||P_0) = \frac{1}{n}\log T$.

From the argument given above, $P_1^* = P_0^* (= P_\lambda^*$ say) and the error exponents:

$$\alpha \approx 2^{-nD(P^*||P_0)}$$

and

$$\beta \approx 2^{-nD(P^*||P_1)}$$

where

$$P_\lambda^* = \frac{P_0^\lambda(x^n) P_1^{1-\lambda}(x^n)}{\sum_{a^n \in \mathcal{X}^n} P_0^\lambda(a^n) P_1^{1-\lambda}(a^n)}.$$

Different choice of threshold T correspond to different λ. Observe that when $\lambda \to 1$, $P^*_\lambda \to P_0$ and when $\lambda \to 0$, $P^*_\lambda \to P_1$, thus giving us the desired tradeoff between false alarm α and miss β probabilities.

If we take a Bayesian approach, the overall probability of error $P_e^{(n)} = \alpha Pr(H_0) + \beta Pr(H_1)$ and define the *best achievable exponent in Bayesian probability of error*,

$$D^* = \lim_{n\to\infty} \min_{A\subseteq\mathcal{X}^n} -\frac{1}{n} P_e^{(n)}.$$

Using the above error exponents for false alarm (type I) and miss (type II) probabilities of error, we have:

Theorem 24.1 (Chernoff Theorem) *The best achievable exponent in Bayesian probability of error*

$$D^* = D(P^*_{\lambda^*}||P_0) = D(P^*_{\lambda^*}||P_1)$$

where λ^* is chosen so that $D(P^*_{\lambda^*}||P_0) = D(P^*_{\lambda^*}||P_1)$. The D^* is commonly known as **Chernoff information**.

Proof: Consider $Pr(H_0), Pr(H_1)$ to be constants not equal to 0 or 1.

$$P_e^{(n)} \approx Pr(H_0)2^{-nD(P^*_\lambda||P_0)} + Pr(H_1)2^{-nD(P^*_\lambda||P_1)} \approx 2^{-n\min(D(P^*_\lambda||P_0), D(P^*_\lambda||P_1))}$$

The right hand side is minimized if λ is such that $D(P^*_\lambda||P_0) = D(P^*_\lambda||P_1)$. ■

Notice that D^* doesn't depend on prior probabilities (unless one of the prior probabilities is vanishingly small), and hence the effect of the prior is washed out for large sample sizes.

If we take a Neyman-Pearson approach instead, and require the probability of false alarm to be fixed (or converging to 0 arbitrarily slowly), what is the best error exponent for the probability of miss?

Theorem 24.2 (Chernoff-Stein's Lemma) *Assume $D(P_0||P_1) < \infty$. For $0 < \epsilon < 1/2$, define*

$$\beta_n^\epsilon = \min_{A\subseteq\mathcal{X}^n, \alpha<\epsilon} \beta$$

Then

$$\lim_{\epsilon\to 0} \lim_{n\to\infty} \frac{1}{n} \log \beta_n^\epsilon = -D(P_0||P_1).$$

Inuitively, if we allow α to be fixed, then $P^*_\lambda = P_0$ (exponent does not decay) and hence $\beta \approx 2^{-nD(P_0||P_1)}$, i.e. we can achieve a faster error exponent on one type of error probability if we allow the other type of error probability to be fixed or decay arbitrarily slowly. For a rigorous proof, see Thomas-Cover Section 11.8.

These results can also be extended to multiple hypothesis testing, and the Chernoff information (exponent of decay of probability of error) is $\min_i D(P_i^*, P_i)$, where P_i^* is the information projection of P_i onto the set of distributions in the complement of A_i (the optimal decision region for hypothesis i). In the NP sense, it is $\min_{i\neq j} D(P_j, P_i)$. See http://ieeexplore.ieee.org/abstract/document/4544997/ for details.

24.2 Cramer Rao Lower Bound

As a last topic, we discuss the Cramer Rao Lower Bound. So far, we had talked about minimax lower bounds for testing and prediction that hold for a class of distributions or for all values of a parameter, but does not let us quantify a performance lower bound for a specific parameter value.

Cramer Rao is a technique for lower bounding the performance, for a specific parameter value, of estimators that are unbiased, though it can be generalized to biased estimators as well. Let $p(x;\theta)$ be a probability density function with continuous parameter $\theta \in \Theta$. Let $X_1, \dots, X_n$ be n i.i.d samples from this distribution, i.e., $X_i \sim p(x;\theta)$. Let $\hat{\theta}(X_1, \dots, X_n)$ be an unbiased estimator of θ, so that $\mathbb{E}\hat{\theta} = \theta$.

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

1. $$\frac{\partial}{\partial\theta}\left[\int \dots \int \hat{\theta}(x_1, \dots, x_n) \prod_{i=1}^{n} p(x_i;\theta)\right] = \int \dots \int \hat{\theta}(x_1, \dots, x_n) \frac{\partial \prod_{i=1}^{n} p(x_i;\theta)}{\partial\theta} dx_1 \dots dx_n \tag{24.1}$$

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

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

then the variance of the unbiased estimator is bounded as:

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

where $I(\theta)$ is the **Fisher Information**. Notice that 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.

We can see that this is an important result as now we are able to bound the variance of unbiased estimators for a specific parameter instead of over a class of parameters.

24.2.1 Proof

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

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

Differentiating both sides w.r.t θ and using Equation 24.1 we get

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

Taking square of both sides,

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

Applying Cauchy-Schwartz inequality ($(\int fg)^2 \leq \int f^2 \cdot \int g^2$), which is applicable under assumption 2 since var is bounded, on RHS assuming the 2 functions to be f and g as shown above,

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

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

24.2.2 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,

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

24.2.3 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 locally for a specific parameter value θ.

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).

24.2.4 Examples

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

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.

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 $cov(\hat{\theta}) \succeq \sigma^2 (A^T A)^{-1}$.

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.

24.2.5 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.