7. Hypothesis Testing and KL Divergence

ECE 830, Spring 2014

Introducing the Kullback-Leibler Divergence

Suppose

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} q(x)$$

and we have two models for $q(x)$:

$$H_0 : q(x) = p_0(x)$$
$$H_1 : q(x) = p_1(x)$$

In past lectures we have seen that the likelihood ratio test (LRT) is optimal, assuming that q is p_0 or p_1. The error probabilities can be computed numerically in many cases.

Error rates

The error probabilities converge to 0 as the number of samples n grows, but numerical calculations do not always yield insight into rate of convergence.

Main result in this lecture

The rate is exponential in n and parameterized the Kullback-Leibler (KL) divergence, which quantifies the differences between the distributions p_0 and p_1.

Our analysis will also give insight into the performance of the LRT when q is neither p_0 nor p_1. This is important since, in practice, p_0 and p_1 may be imperfect models for reality, q in this context. The LRT acts as one would expect in such cases, it picks the model that is closest (in the sense of KL divergence) to q.

To begin our discussion, recall the likelihood ratio is

$$\Lambda = \Lambda(x) = \prod_{i=1}^{n} \frac{p_1(x_i)}{p_0(x_i)}$$

The log likelihood ratio, normalized by dividing by n, is then

$$\hat{\Lambda}_n = \frac{1}{n} \sum_{i=1}^{n} \log \frac{p_1(x_i)}{p_0(x_i)}$$

Note that $\hat{\Lambda}_n$ is itself a random variable, and is in fact a **scaled sum of iid random variables**

$$L_i := \log \frac{p_1(x_i)}{p_0(x_i)}$$

which are independent because the x_i are.

In addition, we know from the strong law of large numbers that for large n, $\hat{\Lambda}_n \xrightarrow{a.s.} \mathbb{E}\left[\hat{\Lambda}_n\right]$, where

KL divergence

The quantity $\int \log \frac{q(x)}{p(x)} q(x) dx$ is known as the *Kullback-Leibler Divergence* of p from q, or the *KL divergence* for short. We use the notation

$$D(q||p) = \int q(x) \log \frac{q(x)}{p(x)} dx$$

for continuous random variables, and

$$D(q||p) = \sum_i q_i \log \frac{q_i}{p_i}$$

for discrete random variables.

The above expression for $\mathbb{E}\left[\hat{\Lambda}_n\right]$ can then be written as

Therefore, for large n, the log likelihood ratio test

$$\hat{\Lambda}_n \underset{H_0}{\overset{H_1}{\gtrless}} \lambda$$

is approximately performing the comparison

$$D(q||p_0) - D(q||p_1) \underset{H_0}{\overset{H_1}{\gtrless}} \lambda$$

since $\hat{\Lambda}_n$ will be close to its mean when n is large. Recall that the minimum probability of error test (assuming equal prior probabilities for the two hypotheses) is obtained by setting $\lambda =$. In this case, we have the test

$$D(q||p_0) \underset{H_0}{\overset{H_1}{\gtrless}} D(q||p_1)$$

For this case, using the LRT is selecting the model that is "closer" to q in the sense of KL divergence.

Example:

Suppose we have the hypotheses

$$H_0 \ : \ X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu_0, \sigma^2)$$
$$H_1 \ : \ X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu_1, \sigma^2)$$

Then we can calculate the KL divergence:

Example: (cont.)

So the KL divergence between two Gaussian distributions with different means and the same variance is just proportional to the squared distance between the two means. In this case, we can see by symmetry that $D(p_1||p_0) = D(p_0||p_1)$, but in general this is not true.

A Key Property

Nonnegativity of KL divergence

$$D(q||p) \geq 0$$

with equality if and only if $q = p$.

To prove this, we will need Jensen's Inequality. If we rearrange the KL divergence formula,

$$\begin{aligned} D(q||p) &= \int q(x) \log \frac{q(x)}{p(x)} dx \\ &= \mathbb{E}_q \left[\log \frac{q(x)}{p(x)} \right] \\ &= -\mathbb{E}_q \left[\log \frac{p(x)}{q(x)} \right] \end{aligned}$$

Now we can use Jensen's inequality, since $-\log z$ is a convex function.

Therefore $D(q||p) \geq 0$.

Bounding the Error Probabilities

The KL divergence also provides a means to bound the error probabilities for a hypothesis test. For this we will need to recall Hoeffding's Inequality:

Hoeffding's Inequality:

If $Z_1, \ldots, Z_n$ are iid and $a \leq Z_i \leq b$, $\forall\ i$, then

$$\mathbb{P}\left(\frac{1}{n}\sum_i Z_i - \mathbb{E}\left[Z\right] > \epsilon\right) \leq e^{-2n\epsilon^2/c^2}$$

and

$$\mathbb{P}\left(\mathbb{E}\left[Z\right] - \frac{1}{n}\sum_i Z_i > \epsilon\right) \leq e^{-2n\epsilon^2/c^2}$$

where $c^2 = (b-a)^2$.

Now suppose that p_0 and p_1 have the same support and that over that support they are both bounded away from zero and from above; i.e. $0 < \alpha \leq p_i(x) \leq \beta < \infty$, $i = 0, 1$. It then follows that

$$\log \frac{\alpha}{\beta} \leq \underbrace{\log \frac{p_1(x_i)}{p_0(x_i)}} \leq \log \frac{\beta}{\alpha}$$

Thus L_i is bounded, and $\hat{\Lambda}_n$ is a sum of iid bounded random variables. This allows us to use Hoeffding's Inequality.

Now, consider the hypothesis test $\hat{\Lambda}_n \underset{H_0}{\overset{H_1}{\gtrless}} 0$. We will now assume that the data $X_1, \ldots, X_n \overset{iid}{\sim} q$, with q either p_0 or p_1. We can write the probability of false alarm as

$$\begin{aligned} P_{FA} &= \mathbb{P}\left(\hat{\Lambda}_n > 0 | H_0\right) \\ &= \mathbb{P}\left(\hat{\Lambda}_n - \mathbb{E}\left[\hat{\Lambda}_n | H_0\right] > -\mathbb{E}\left[\hat{\Lambda}_n | H_0\right] \mid H_0\right) \end{aligned}$$

The quantity $-\mathbb{E}\left[\hat{\Lambda}_n | H_0\right]$ will be the ϵ in Hoeffding's inequality. We can re-express it as

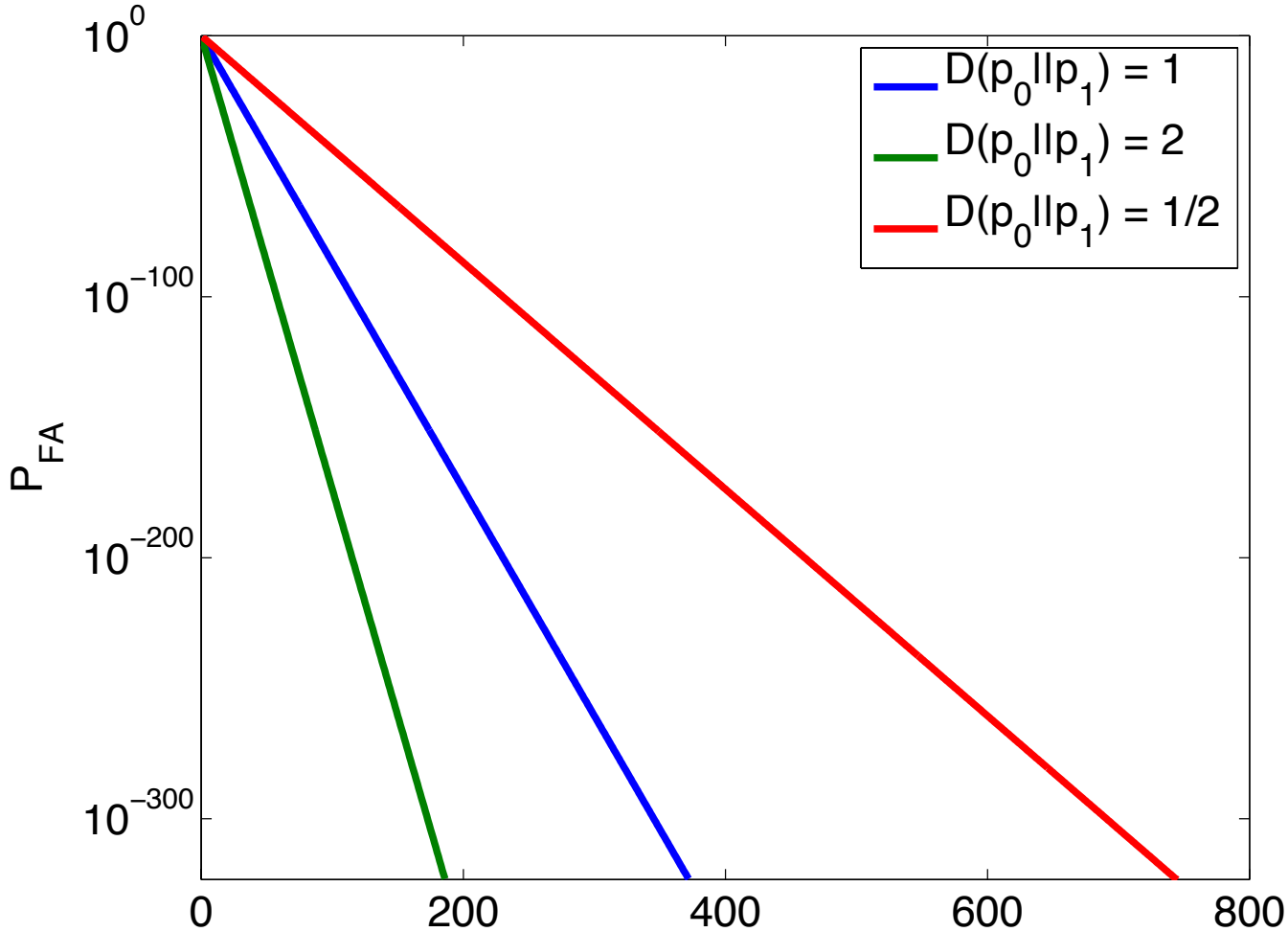
$$\begin{aligned} \mathbb{E}_{p_0}\left[\hat{\Lambda}_n | H_0\right] &= \int p_0(x) \log \frac{p_1(x)}{p_0(x)} dx \\ &= -\int p_0(x) \log \frac{p_0(x)}{p_1(x)} dx \\ &= -D(p_0 || p_1) \end{aligned}$$

Finally applying Hoeffding's inequality, we get

$$\begin{aligned} P_{FA} &= \mathbb{P}\left(\hat{\Lambda}_n - (-D(p_0||p_1)) > D(p_0||p_1) \mid H_0\right) \\ &\leq e^{-2nD^2(p_0||p_1)/c^2} \end{aligned}$$

with $c^2 = \left(\log\frac{\beta}{\alpha} - \log\frac{\alpha}{\beta}\right)$.
Thus the probability of false alarm error is bounded using the KL divergence $D(p_0||p_1)$. **As n or $D(p_0||p_1)$ increase, the error decreases exponentially.** The bound for the probability of miss, the other type of error, can be found in a similar fashion:

$$\begin{aligned} P_M &= \mathbb{P}\left(\hat{\Lambda}_n < 0 \mid H_1\right) \\ &= \mathbb{P}\left(\hat{\Lambda}_n - D(p_1||p_0) < -D(p_1||p_0) \mid H_1\right) \\ &= \mathbb{P}\left(D(p_1||p_0) - \hat{\Lambda}_n > D(p_1||p_0) \mid H_1\right) \\ &\leq e^{-2nD^2(p_1||p_0)/c^2} \end{aligned}$$



The slope of each line is proportional to $D(p_0\|p_1)$. For this reason, $D(p_0\|p_1)$ is referred to as the **exponential rate**.