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

Lecture 22: Error exponents in hypothesis testing, GLRT

Lecturer: Aarti Singh *Scribe: Aarti Singh*

22.1 Recap

In last lecture, we discussed Sanov's theorem which provides a general bound on large deviations, i.e. probability of atypical sequences under the true distribution.

Theorem 22.1 (Sanov's Theorem) *Let $X_1, \ldots, X_n \overset{iid}{\sim} Q$. Let $E \subseteq \mathcal{P}$ be a set of probability distributions. Then*

$$Q^n(E) = \sum_{x^n : P_{x^n} \in E \cap \mathcal{P}_n} Q^n(x^n) \leq (n+1)^{|\mathcal{X}|}\ 2^{-nD(P^*||Q)}$$

where $P^* = \arg\min_{P \in E} D(P||Q)$.
If, in addition, set E is the closure of its interior, then

$$\frac{1}{n} \log Q^n(E) \to -D(P^*||Q).$$

Remark 1: The theorem says that the probability of set E under a distribution Q is the same as the probability of the type P^* in E that is closest to Q (in terms of KL distance) up to first order in exponent.

Remark 2: The polynomial term in the bound can be dropped if E is a convex set of distributions.

Some specific examples of application of Sanov's theorem are as follows:

1. Suppose $X_1, \ldots, X_n \overset{iid}{\sim} Bernoulli(1/3) = Q$ and we want to find the probability that $\frac{1}{n}\sum_{i=1}^n X_i \geq 3/4$.

   $$Q^n(\{x^n : \sum_{a \in \{0,1\}} a P_{x^n}(a) \geq 3/4\}) = Q^n(E = \{P : P(1) \geq 3/4\}) \leq (n+1)^{|\mathcal{X}|}\ 2^{-nD(P^*||Q)}$$

   where $P^* = \arg\min_{P \in E} D(P||Q) = \arg\min_{P: P(1) \geq 3/4} D(P||Q) = (1/4, 3/4)$ since type $(1/4, 3/4)$ in E is closest to the true distribution $Q = (2/3, 1/3)$. Thus,

   $$Q^n(E) \approx 2^{-nD((1/4,3/4)\ ||\ (2/3,1/3))}$$

   asymptotically.

2. Let $Q(X, Y)$ be some joint distribution. Suppose $(X^n, Y^n) = (X_1, Y_1), \ldots, (X_n, Y_n) \overset{iid}{\sim} Q_0(X, Y) = Q(X)Q(Y)$, where $Q(X)$ and $Q(Y)$ are the marginal distributions corresponding to the joint distribution $Q(X, Y)$. In other words, X_i and Y_i are independent. We are interested in the probability that (X^n, Y^n) appear to be depdendent or jointly distributed according to $Q(X, Y)$. In last lecture, we saw that using Sanov's theorem we get:

   $$Q_0^n(E) \approx 2^{-nD(Q(X,Y)||Q(X)Q(Y))} = 2^{-nI(X,Y)}$$

Notice that this example essentially corresponds to an **independence test** where

$$H_0 : X \perp Y$$

$$H_1 : X \not\perp Y$$

and Sanov's theorem tells us that the probability of false alarm (type I error) asymptotically scales as $2^{-nI(X,Y)}$ for this test. Independence tests are useful in many problems, e.g. in communication, we used joint typicality decoding to decode the channel output to a codeword with which it was dependent. In machine learning, independnece tests are used for feature selection, i.e. deciding whether or not to discard a feature X based on if the label Y is dependent on it or not. (Conditional) independence tests are used for in causal inference and learning graphical models where an edge between two nodes X, Y is absent if they are independent conditioned on some (set of) variables Z.

Today, we will see the use of Sanov's theorem in characterizing the error exponent of general hypothesis testing problems. But first, we present a proof of Sanov's theorem based on results we proved last time for the method of types.

22.2 Proof of Sanov's Theorem

Recall that the type class of P is the set of all sequences with type P, i.e. $T(P) = \{x^n : P_{x^n} = P\}$ and the probability of a type class $T(P)$ under Q, $Q^n(T(P)) \leq 2^{-nD(P||Q)}$ and $Q^n(T(P)) \geq \frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-nD(P_n||Q)}$. We use these results to establish Sanov's theorem.

Upper bound:

$$\begin{aligned} Q^n(E) &= \sum_{P \in E \cap \mathcal{P}_n} Q^n(T(P)) \leq \sum_{P \in E \cap \mathcal{P}_n} 2^{-nD(P||Q)} \leq \sum_{P \in E \cap \mathcal{P}_n} \max_{P \in E \cap \mathcal{P}_n} 2^{-nD(P||Q)} \\ &\leq \sum_{P \in E \cap \mathcal{P}_n} 2^{-n \min_{P \in E} D(P||Q)} \leq (n+1)^{|\mathcal{X}|} \, 2^{-nD(P^*||Q)} \end{aligned}$$

The last step follows since the total number of types $|\mathcal{P}_n| \leq (n+1)^{|\mathcal{X}|}$. This also implies that

$$\limsup \frac{1}{n} \log Q^n(E) \to -D(P^*||Q)$$

Lower Bound:

If E is the closure of its interior, then it implies that E is non-empty. Also observe that $\cup_n \mathcal{P}_n$, the set of all types for all n, is dense in all distributions. These two facts imply that $E \cap \mathcal{P}_n$ is also non-empty for large enough n and that we can find a type $P_n \in E \cap \mathcal{P}_n$ s.t. $D(P_n||Q) \to D(P^*||Q)$. Now

$$Q^n(E) \geq Q^n(T(P_n)) \geq \frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-nD(P_n||Q)}$$

This implies that

$$\liminf \frac{1}{n} \log Q^n(E) \to -D(P^*||Q)$$

which completes the proof.

22.3 Error Exponents in Hypothesis Testing

22.3.1 Hypothesis testing: Neyman-Pearson and Bayesian

We now use Sanov's theorem to establish error exponents in hypothesis testing. We consider the following binary hypothesis testing problem: Let $X_1, \ldots, X_n \overset{iid}{\sim} Q$.

$$H_0 : Q = P_0$$

$$H_1 : Q = P_1$$

The test is specified as a decision function $g(X_1, \ldots, X_n) \in \{0, 1\}$, which maps the sequence X^n to one of the two hypothesis. The corresponding decision region A is given by the set of sequences that are mapped to hypothesis H_0, i.e. $A = \{x^n : g(x^n) = 0\}$. The test is associated with two types of error:

$$\begin{aligned} \text{Probability of false alarm/type I error} \qquad & \alpha = Pr(g = 1|H_0) = P_0(A^c) \\ \text{Probability of miss/type II error} \qquad & \beta = Pr(g = 0|H_1) = P_1(A) \end{aligned}$$

There are two approaches to hypothesis testing based on the kind of error control desired:

- **Neyman-Pearson approach:** Minimize the probability of miss (type II error) subject to a desired control on the probability of false alarm (type I error): $\min \beta \ s.t. \ \alpha \leq \epsilon$
- **Bayesian approach:** If we have some prior belief over the probabilities of the two hypotheses, then we minimize the expected probability of error: $\alpha Pr(H_0) + \beta Pr(H_1)$

The following lemma tells us the form of the optimal test under the Neyman-Pearson approach:

Theorem 22.2 (Neyman-Pearson Lemma) *For a threshold $T \geq 0$, define the decision region corresponding to a likelihood ratio test $A(T) = \left\{ \frac{P_0(x^n)}{P_1(x^n)} > T \right\}$. Let $\alpha^* = P_0(A^c(T))$ be the false alarm (type I error probability) and let $\beta^* = P_1(A(T))$ be the miss (type II error probability) of this test. Let B be any other decision region with associated false alarm and miss probabilities α and β. Then if $\alpha < \alpha^*$, then $\beta > \beta^*$.*

Proof: First, we will show that for all sequences $x^n \in \mathcal{X}^n$,

$$[1_A(x^n) - 1_B(x^n)](P_0(x^n) - TP_1(x^n)) \geq 0$$

where 1_S denotes the indicator function of the set S, i.e. $1_S(x^n) = 1$ if $x^n \in S$ and 0 if $x^n \notin S$. To see this, consider two cases: i) $x^n \in A$, then first term is positive (≥ 0) and by definition of $A(T)$, the second term is positive as well. ii) $x_n \notin A$, then first term is negative (≤ 0) and by definition of $A(T)$, the second term is negative as well. Thus, the product of the two terms is positive in both cases.

Summing over all sequences and expanding out the terms

$$\begin{aligned} 0 &\leq \sum_{x^n} [1_A(x^n)P_0(x^n) - 1_B(x^n)P_0(x^n) - T1_A(x^n)P_1(x^n) + T1_B(x^n)P_1(x^n)] \\ &= P_0(A(T)) - P_0(B) - TP_1(A(T)) + TP_1(B) \\ &= 1 - \alpha^* - (1 - \alpha) - T\beta^* + T\beta \\ &= \alpha - \alpha^* + T(\beta - \beta^*) \end{aligned}$$

This implies that if the first term is negative ($\alpha < \alpha^*$), then second term has to be positive ($\beta > \beta^*$). ■

Thus, the likelihood ratio test is the optimal test which achieves the best miss probability for a given probability of false alarm. The threshold T is chosen to meet the desired probability of false alarm α.

You should convince yourself that the optimal Bayesian test is based on the ratio of aposterior probabilities (instead of likelihoods) - maybe a HW problem

$$\frac{P_0(x^n)Pr(H_0)}{P_1(x^n)Pr(H_1)} > 1 \quad \equiv \quad \frac{P_0(x^n)}{P_1(x^n)} > \frac{Pr(H_1)}{Pr(H_0)}$$

22.3.2 Information-theoretic interpretation

Lets re-write the log likelihood ratio test in terms of information theoretic quantities.

$$\begin{aligned}
\text{Log likelihood ratio} &= \log \frac{P_0(x^n)}{P_1(x^n)} = \sum_{i=1}^{n} \log \frac{P_0(x_i)}{P_1(x_i)} \\
&= \sum_{a \in \mathcal{X}} n P_{x^n}(a) \log \frac{P_0(a)}{P_1(a)} = \sum_{a \in \mathcal{X}} n P_{x^n}(a) \log \frac{P_{x^n}(a)}{P_1(a)} \cdot \frac{P_0(a)}{P_{x^n}(a)} \\
&= n[D(P_{x^n}||P_1) - D(P_{x^n}||P_0)]
\end{aligned}$$

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 22.1.

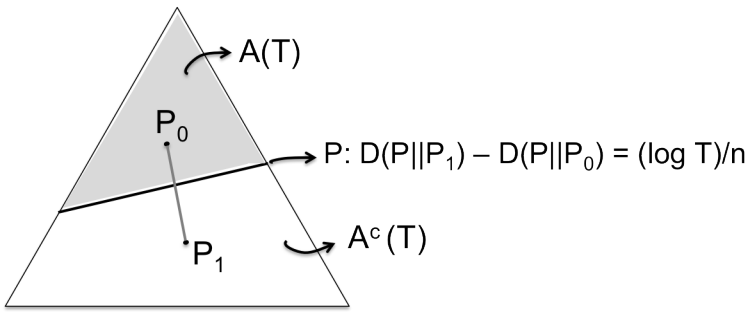

Figure 22.1: The decision region corresponding to a likelihood ratio test is demarcated by boundary that is parallel to the perpendicular bisector of the line joining the distributions under the two hypotheses P_0 and P_1.

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

Let us evaluate the form of P_1^* (and P_0^*). Notice from Figure 22.1 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 $P(x^n)$:

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

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 22.3 (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 $Pr(H_0), Pr(H_1)$ 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 probabiltiies is 0), 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 22.4 (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.

22.4 GLRT (Generalized Likelihood Ratio Test)

So far we have assumed that the distributions under H_0 and H_1 are perfectly known and the likelihood or aposteriori ratio is computable. However, in practice this is not the case. Let us look at a simple hypothesis testing problem of normal means and some practical tests for that setting.

In the normal means problem, the two hypothesis of interest are:

$$H_0 : X \sim \mathcal{N}(0, \sigma^2)$$

$$H_1 : X \sim \mathcal{N}(\mu, \sigma^2)$$

e.g. in a classification problem, if we assume that class conditional densities are Gaussian, i.e. $p(x|Y) \sim \mathcal{N}(\mu_y, \sigma^2)$, then the classification problem is essentially the normal means hypothesis testing problem stated above. The likelihood ratio test in this setting corresponds to the test statistic $x\mu$ (or $x^T\mu$ in multi-variate case). This is known as a **matched filter** since we are essentially testing matching the observation x with the known signal μ.

If μ is not known, then we have a composite hypothesis testing problem:

$$\begin{aligned} H_0 :&\quad X \sim \mathcal{N}(0, \sigma^2) \\ H_1 :&\quad X \sim \mathcal{N}(\mu, \sigma^2), \qquad \mu > 0 \end{aligned}$$

This is particulary the case in problems such as anomaly detection where the abnormal case corresponds to some unknown activation $\mu > 0$. In this case, there are two approaches: (1) Bayesian - here we assume a prior on the unknown parameter μ and consider the Bayesian LRT:

$$\frac{P_0(x)}{\mathbb{E}_\mu[P_{1,\mu}(x)]} > T$$

or (2) Generalized likelihood ratio test (GLRT):

$$\frac{P_0(x)}{\max_\mu P_{1,\mu}(x)} = \frac{P_0(x)}{P_{1,\hat{\mu}_{mle}}(x)} > T$$

This is essentially a plug-in method where we plug-in the maximum likelihood estimate of the parameter under the alternate hypothesis. For the normal means problem, the GLRT simplifies to the test statistic x^2 (or $x^T x = \|x\|^2$ in multi-variate setting) since the MLE of the mean is simply x (if there is one sample or $\bar{x} = \sum_{i=1}^n x_i/n$ if there are n samples). This is known as the **energy detector** since $\|x\|^2$ is the total energy in the signal.

Remark: While the GLRT is a natural solution to unknown parameters, it may not be optimal in all settings. For example, consider the high-dimensional setting where x is a sparse d-dimensional vector with only k non-zeros. Then the energy detector will have very poor performance because the non-zero signal components are averaged with noise energy in the remaining components. On the other hand, another simple test based on the max statistic $\max_i x_i = \|x\|_\infty$ can perform much better. In fact, it can be shown that that the energy detector asymptotically drives the false alarm and miss probabilities to zero only if $\mu \gg \sqrt{\sigma^2 d}$. On the other hand, the max detector works even if $\mu > \sqrt{2\sigma^2 \log d}$ since the maximum of d iid draws from a standard Gaussian distribution is $< \sqrt{2\sigma^2 \log d}$ with probability $\to 1$ for large d.