Lecture 21: Hypothesis Testing, Method of Types and Large Deviation

Lecturer: Aarti Singh *Scribes: Yu Cai*

21.1 Hypothesis Testing

One of the standard problems in statistics is to decide between two alternative explanations for the data observed. For example, in medical testing, one may wish to test whether or not a new drug is effective. Similarly, a sequence of coin tosses may reveal whether or not the coin is biased. These problems are examples of the general hypothesis-testing problem. In the simplest case, we have to decide between two i.i.d. distributions. For example, the transmitter sends the information bits by bits in communication systems. There are two possible cases for each transmission: one is that bit 0 is sent (noted as event H0) and the other is that bit 1 is sent (noted as event H1). In the receiver side, the bit y is be received as either 0 or 1. Based on the y bit received, we can make a hypothesis whether the event H0 happens (bit 0 was sent at the transmitter) or the event H1 happens (i.e. bit 1 was sent at the transmitter). Of course, we may make mis-judgement, such as we decode that bit 0 was sent but actually bit 1 was sent. We need to make the probability of error in hypothesis testing as low as possible.

To be general, let $X_1, X_2, ..., X_n$ be $\overset{i.i.d}{\sim}$ Q(x). We can consider two hypothesis:

- H_0: Q = P_0. (null hypothesis)
- H_1: Q = P_1. (alternative hypothesis)

Consider the general decision function $g(x_1, x_2, ..., x_n)$, where $x_i \in \{0, 1\}$. When $g(x_1, x_2, ..., x_n) = 0$ means that H_0 is accepted and $g(x_1, x_2, ..., x_n) = 1$ means that H_1 is accepted. Since the function takes on only two values, the test can be specified by specifying the set A over which $g(x_1, x_2, ..., x_n)$ is 0. The complement of this set is the set where $g(x_1, x_2, ..., x_n)$ has the value 1. The set A knwon as the *decision region* can be expressed as

$$A = \{x^n : g(x^n) = 0\}.$$

There are two probabilities of error as follows:

1. Type I (False Alarm):
$$\alpha_n = Pr(g(x_1, x_2, ..., x_n) = 1|\text{event } H_0 \text{ is true})$$
2. Type II (Miss):
$$\beta_n = Pr(g(x_1, x_2, ..., x_n) = 0|\text{event } H_1 \text{ is true})$$

In general, we wish to minimize the probabilities of both false alarm and miss. But there is a tradeoff. Thus, we minimize one of the probabilities of error subject to a constraint on the other probability of error. The best achievable error component in the probability of error for this problem is given by the Chernoff-Stein lemma. There are two types of approaches to hypothesis testing based on the kind of error control needed:

1. Neyman-Pearson approach: To minimize the probability of miss given an acceptable probability of false alarm. It can be expressed as $\min_g \beta$ (such that $\alpha \leq \epsilon$).
2. Bayesian approach : The goal is to minimize the expected probability of both false alarm and miss, where we assume a prior distribution on the two hypotheses $P(H_0)$ and $P(H_1)$. It can be expressed as $\min_g \beta_n P(H_1) + \alpha_n P(H_0)$.

Theorem 21.1 (Neyman-Pearson lemma) *Let $X_1, X_2, ..., X_n$ be drawn i.i.d according to probability mass function Q. Consider the decision problem corresponding to hypothesis $H_0 : Q = P_0$ vs $H_1 : Q = P_1$. For T $\geq$ 0, define a region*

$$A_n(T) = \{x^n : \frac{P_0(x_1, x_2, ..., x_n)}{P_1(x_1, x_2, ..., x_n)} \geq T\}$$

Let

$$\alpha^* = P_0(A_n^c(T)) \text{ (False Alarm)}$$

$$\beta^* = P_1(A_n(T)) \text{ (Miss)}$$

be the corresponding probabilities of error corresponding to decision region A_n. Let B_n be any other decision region with associated probabilities of α and β. Then, if $\alpha < \alpha^*$ then $\beta > \beta^*$, and if $\alpha = \alpha^*$ then $\beta >= \beta^*$.

The proof of this theorem will be explained in the next lecture.

Note: In the Bayesian setting, we can similarly construct the test with the optimal Bayesian error:

$$A_n = \{x^n : \frac{P_0(x^n)P(H_0)}{P_1(x^n)P(H_1)} \geq 1\}$$

To study how the proability 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). But before we get to that, we need to understand the method of types.

21.2 Method of types

In a previous lecture, we have introduced the AEP concept for discrete random variables, which focuses attention on a small subset of typical sequences. In this section, we will introduce the concept of method of types, which is a more powerful procedure in which we consider set of sequences that have the same empirical distribution. Based on this restriction, we will derive strong bounds on the number of sequences with a particular empirical distribution.

Type: The type P_{x^n} (or empirical probability distribution) of a sequence x_1, x_2, ..., x_n is the relative proportion of occurrences of each symbol $a \in \mathcal{X}$ (i.e. $P_{x^n}(a) = \frac{N(a,x^n)}{n}$ for all $a \in X$, where $N(a, x^n)$ is the number of times the symbol a occurs in the sequence $x^n \in \mathcal{X}^n$). The type of a sequence x^n is denoted as P_{x^n} and it is a probability mass function on $\mathcal{X}$.

Set of types: Let $\mathcal{P}_n$ denote the set of types with denominator n. For example, if $\mathcal{X}$=$\{0, 1\}$, the set of possible types with denominator n is

$$P_n = \{(P(0), P(1)) : (\frac{0}{n}, \frac{n}{n}), (\frac{1}{n}, \frac{n-1}{n}), ..., (\frac{n}{n}, \frac{0}{n})\}$$

Type class: If P $\in P_n$, the set of sequences of length n and type P is called the type class of P, denoted as T(P):

$$T(P) = \{x^n \in \mathcal{X}_n : P_{x^n} = P\}.$$

Theorem 21.2

$$|\mathcal{P}_n| \leq (n+1)^{|\mathcal{X}|}$$

Proof: There are $|\ \mathcal{X}\ |$ components in the vector that specifies P_{x^n}. The numerator in each component can take on only $n+1$ values. So there are at most $(n+1)^{|\mathcal{X}|}$ choices for the type vector. Of course, these choices are not independent, but this is sufficient good upper bound. ■

Theorem 21.3 *If $x^n = (x_1, x_2, ..., x_n)$ are drawn i.i.d according to Q(x), the probability of x^n depends only on its type and is given by*

$$Q^n(x^n) = 2^{-n(H(P_{x^n}) + D(P_{x^n}||Q))}$$

Proof:

$$\begin{aligned}
Q^n(x^n) &= \prod_{i=1}^{n} Q(x_i) = \prod_{a \in \mathcal{X}} Q(a)^{N(a,x^n)} \\
&= \prod_{a \in \mathcal{X}} Q(a)^{nP_{x^n}(a)} = \prod_{a \in \mathcal{X}} 2^{nP_{x^n}(a)\log Q(a)} \\
&= \prod_{a \in \mathcal{X}} 2^{n(P_{x^n}(a)\log Q(a) - P_{x^n}(a)\log P_{x^n}(a) + P_{x^n}(a)\log P_{x^n}(a))} \\
&= 2^{n\sum_{a\in\mathcal{X}}(-P_{x^n}(a)\log\frac{P_{x^n}(a)}{Q(a)}) + P_{x^n}(a)\log P_{x^n}(a))} \\
&= 2^{-n(D(P_{x^n}||Q) + H(P_{x^n}))}
\end{aligned}$$

■

Based on the above theorem, we can easily get the following results. If x^n is in the type class of Q, that is $x^n \in T(Q)$ then

$$Q^n(x^n) = 2^{-nH(Q)}.$$

Theorem 21.4 *(Size of a type class T(P)) For any type $P \in P_n$,*

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{nH(P)} \leq |T(P)| \leq 2^{nH(P)}$$

This theorem gives an estimate of the size of a type class T(P).

The upper bound follows by considering $P^n(T(P)) \leq 1$ and lower bounding this by the size of the type class and lower bound on the probability of sequences in the type class. The lower bound is a bit more involved, see Cover-Thomas proof of Thm 11.1.3.

Theorem 21.5 *(Probability of type class) For any $P \in \mathcal{P}_n$ and any distribution Q, the probability of the type class T(P) under Q^n is $2^{-nD(P||Q)}$ for first order in the exponent. More precisely,*

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-nD(P||Q)} \leq Q^n(T(P)) \leq 2^{-nD(P||Q)}$$

Proof:

$$\begin{aligned} Q^n(T(P)) &= \sum_{x^n \in T(P)} Q^n(x^n) \\ &= \sum_{x^n \in T(P)} 2^{-n(D(P||Q)+H(P))} \\ &= |T(P)| 2^{-n(D(P||Q)+H(P))} \end{aligned}$$

Using the bounds on $|T(P)|$ derived in theorem 21.4, we have the following results

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-nD(P||Q)} \leq Q^n(T(P)) \leq 2^{-nD(P||Q)}$$

■

21.3 Large Deviation Theory

The subject of large deviation theory can be illustrated by one example as follows. The event that $\frac{1}{n}\sum X_i$ is near $\frac{1}{3}$ if $X_1, X_2, ..., X_n$ are drawn i.i.d Bernoulli($\frac{1}{3}$) is a small deviation. But the probability that $\frac{1}{n}\sum X_i$ is greater than $\frac{3}{4}$ is a large deviation. We will show that the large deviation probability is exponentially small. Note that $\frac{1}{n}\sum X_i = \frac{3}{4}$ is equivalent to $P_{x^n} = (\frac{1}{4}, \frac{3}{4})$. Recall that the probability of a sequence $\{X_i\}_{i=1}^n$ depends on its type P_{x^n}. Amongst sequences with $\frac{1}{n}\sum X_i \geq \frac{3}{4}$, the closest type to the true distribution is $(\frac{1}{4}, \frac{3}{4})$, and we will show that the probability of the large deviation will turn out to be around $2^{-nD((\frac{1}{4}),(\frac{3}{4}))\ ||\ ((\frac{2}{3},\frac{1}{3})))}$.

Theorem 21.6 (Sanov theorem) *Let $X_1, X_2, ..., X_n$ be i.i.d Q(x) distribution. Let $E \subseteq P$ be a set of probability distributions. Then*

$$Q^n(E) = Q^n(E \cap P_n) \leq (n+1)^{|\mathcal{X}|} 2^{-nD(P^*||Q)}$$

where

$$P^* = \arg\min_{P \in E} D(P||Q)$$

is the distribution in E that is closest to Q in relative entropy, i.e. the Information-projection of Q onto E. If in addition, the set E is the closure of its interior, then

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

The proof will be disucssed in the next lecture.

In the following sections, we will show two examples of using the Sanov's theorem.
Example 1: Suppose that we wish to find $Pr\{\frac{1}{n}\sum_{i=1}^n g_j(X_i) \geq \alpha_j, j = 1, 2, ..., k\}$.. Since $\frac{1}{n}\sum_{i=1}^n g_j(X_i) = \sum_{a \in \mathcal{X}} P_{x^n}(a) g_j(a)$, the set E is defined as

$$E = \{P : \sum_a P(a) g_j(a) \geq \alpha_j, j = 1, 2, ..., k\}$$

To find the closest distribution in E to Q, we need to minimize $D(P||Q)$ subject to the constraints. Using Lagrange multipliers, we construct the functional

$$J(P) = \sum_x P(x) \log \frac{P(x)}{Q(x)} + \sum_j \lambda_j \sum_x P(x) g_j(x) + v \sum_x P(x)$$

We can differentiate and setting the derivative equal to zero, we calculate the closest distribution to Q to be of the form

$$P^*(x) = \frac{Q(x)e^{\sum_j \lambda_j g_j(x)}}{\sum_{a \in \mathcal{X}} Q(a)e^{\sum_j \lambda_j g_j(a)}}$$

where the constants λ_j are chosen to satisfy the constraints. Note that if Q is uniform, P^* is the maximum entropy distribution. Thus, $Q^n(E)$ asymptotically follows the distribution as $2^{-nD(P^*||Q)}$ by Sanov's theorem.

For the Bernoulli example mentioned above, there is only one g and the constraint set corresponds to $g(a) = a$. Since $Q \sim Bernoulli(2/3, 1/3)$, we have $Q(x) = (2/3)^{1-x}(1/3)^x = (2/3) * (1/2)^x$

$$P^*(x) = \frac{\frac{2}{3}\left(\frac{1}{2}\right)^x e^{\lambda x}}{\sum_{a \in \{0,1\}} \frac{2}{3}\left(\frac{1}{2}\right)^a e^{\lambda a}} = \frac{\left(\frac{1}{2}\right)^x e^{\lambda x}}{1 + \left(\frac{1}{2}\right) e^{\lambda}}$$

For P^* to satisfy the constraint, we must have λ such that $\sum_a aP^*(a) = 3/4$, or equivalently $P^*(1) = 3/4$. This implies that $e^{\lambda} = 6$. This yields

$$P^*(x) = \frac{3^x}{4}$$

i.e. $P^* = (1/4, 3/4)$, which is precisely the distribution which meets the observation constraint $\frac{1}{n}X_i \geq 3/4$ and is closest to the true distribution. Thus, the probability that $\frac{1}{n}X_i \geq 3/4$ when $X_i \sim Q = Bernoulli(1/3)$, is asymptotically $2^{-nD((1/4,3/4)\ ||\ (2/3,1/3))}$ by Sanov's theorem.

Example 2 (Mutual dependence): Let $Q(x, y)$ be a given joint distribution and let $Q_0(x, y) = Q(x)Q(y)$ be the associated product distribution formed from the marginals of Q. We wish to know the likelihood that a sample drawn according to Q_0 will appear to be jointly distributed according to Q. Accordingly, let (X_i, Y_i) be i.i.d and $Q_0(x, y) = Q(x)Q(y)$. We define (x^n, y^n) to be jointly typical with respect to a joint distribution $Q(x, y)$ if the sample entropies are close to their true values as follows:

$$|-\frac{1}{n}\log Q(x^n) - H(X)| \leq \epsilon$$

$$|-\frac{1}{n}\log Q(y^n) - H(Y)| \leq \epsilon$$

$$|-\frac{1}{n}\log Q(x^n, y^n)) - H(X, Y)| \leq \epsilon$$

Thus, (x^n, y^n) are jointly typical with respect to $Q(x, y)$ if the type $P_{x^n,y^n} \in E \subseteq \mathcal{P}_n(X, Y)$, where

$$E = \{P(x, y) : |-\sum_{x,y} P(x, y)\log Q(x) - H(X)| \leq \epsilon,$$

$$|-\sum_{x,y} P(x, y)\log Q(y) - H(Y)| \leq \epsilon,$$

$$|-\sum_{x,y} P(x, y)\log Q(x, y) - H(X, Y)| \leq \epsilon\}$$

Using Sanov theorem, the probability is

$$Q_0^n(E) = 2^{-nD(P^*||Q_0)}$$

where P^* is the distribution satisfying the constraints that is closest to Q_0 in relative entropy. In this case, as $\epsilon \to 0$, it can be verified that P^* is the joint distribution Q, and Q_0 is the product distribution formed from the marginals of Q. So that the probability is $2^{-nD(Q(x,y)||Q(x)Q(y))} = 2^{-nI(X;Y)}$. Note that this is the same results as we can get in previous lectures when applying AEP. To find

$$P^* = \arg\min_{P \in E} D(P||Q_0)$$

we use Lagrange multipliers and construct the Lagrangian function (for $\epsilon = 0$)

$$D(P||Q_0) + \lambda_1 \sum_{x,y} P(x,y) \log Q(x) + \lambda_2 \sum_{x,y} P(x,y) \log Q(y) + \lambda_3 \sum_{x,y} P(x,y) \log Q(x,y) + \lambda_4 \sum_{x,y} P(x,y)$$

Taking derivative wrt $P(x,y)$ and setting it equal to zero, we can calculate the closest distribution as

$$P^* = Q_0 e^{\lambda_1 \log Q(x) + \lambda_2 \log Q(y) + \lambda_3 \log Q(x,y) + \lambda_4}$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are chosen to satisfy the constraints:

$$\begin{aligned}
\sum_{x,y} P^*(x,y) \log Q(x) &= -H(X) = \sum_{x} Q(x) \log Q(x) \\
\sum_{x,y} P^*(x,y) \log Q(y) &= -H(Y) = \sum_{y} Q(y) \log Q(y) \\
\sum_{x,y} P^*(x,y) \log Q(x,y) &= -H(X,Y) = \sum_{x,y} Q(x,y) \log Q(x,y)
\end{aligned}$$

It is easy to check that all constraints are satisfied if $P^* = Q(x,y)$.