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

Lecture 23: Dec 5

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.*

In this lecture, we will look at information-theoretic tools to bound probability of large deviations (and as a consequence concentration inequalities) via Sanov's Theorem. We also use such bounds for error exponent in hypothesis testing.

23.1 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. Amongst distributions P with $E_P[X_i] \geq \frac{3}{4}$, the closest 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 $2^{-nD((\frac{1}{4}),(\frac{3}{4}))\ ||\ ((\frac{2}{3},\frac{1}{3})))}$.

Essentially, large deviation theory is same as concentration inequalities you might have seen in other courses, but it can bound probability of general large deviation events. The key tool for large deviation is Sanov's theorem:

Theorem 23.1 (Sanov theorem) *Let $X_1, X_2, ..., X_n$ be i.i.d Q(x) distribution taking finitely many values, and denote its empirical distribution as P_n. Let E be a set of probability distributions. Then*

$$Q^n(P_n \in E) \leq (n+1)^{|\chi|} 2^{-nD(P^*||Q)}$$

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

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

Remark 1: The theorem says that the probability of set E under a distribution Q is the same as the probability of the distribution 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.

Remark 3: The convergence result (last expression) holds for continuous random variables too, under mild assumptions (see e.g. General form of Sanov's theorem at https://blogs.princeton.edu/sas/2013/10/10/lecture-3-sanovs-theorem/).

The proof for finitely many outcomes will be discussed later and will use the method of types.

Lets consider some 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 $\frac{1}{n}\sum_{i=1}^{n} g_j(X_i) = \sum_{a\in\mathcal{X}} P_n(a)g_j(a)$, 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. This is precisely how we computed information projection earlier. 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

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

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

For the Bernoulli example mentioned earlier, 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 $\frac{1}{n}\sum_{i=1}^{n} X_i \geq 3/4$ and is closest to the true distribution. Thus, the probability that $\frac{1}{n}\sum_{i=1}^{n} X_i \geq 3/4$ when $X_i \sim Q = Bernoulli(1/3)$, is asymptotically $2^{-nD((1/4,3/4)\ ||\ (2/3,1/3))}$ by Sanov's theorem.

In general, if the true distribution $Q \sim Bernoulli(q)$ and we want to bound the deviation that we observe pn or more heads, then Sanov's theorem states that the probability of error is asymptotically $2^{-nD(p,q)}$ which using Pinsker's inequality can be bounded by $2^{-2n(p-q)^2}$ giving us the Chernoff-Hoeffding Bound.

Example 2 (Independence testing): Testing if two variables are independent or not is an important problem that comes up in many applications of machine learning as well as signal processing. In machine learning, independence 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 causal inference and learning graphical models. Also recall that when proving channel coding theorem, we were testing whether a received codeword y^n is jointly typical with a candidate sent codeword x^n. The probability that two independent sequences (x^n, y^n) (x^n being a codeword other than what was sent when y^n was received) actually appear as dependent was bounded asymptotically as $2^{-nI(X,Y)}$. This is essentially the independence testing problem and Sanov's theorem allows us to recover the same result as follows.

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 probability that a sample drawn according to Q_0 will appear to be jointly distributed according to Q. Accordingly, let $\{x_i, y_i\}_{i=1}^{n}$ be i.i.d. drawn from

$Q_0(x,y) = Q(x)Q(y)$. 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:

$$|-\frac{1}{n}\sum_{i=1}^{n}\log Q(x_i) - H(X)| \leq \epsilon$$

$$|-\frac{1}{n}\sum_{i=1}^{n}\log Q(y_i) - H(Y)| \leq \epsilon$$

$$|-\frac{1}{n}\sum_{i=1}^{n}\log Q(x_i,y_i) - H(X,Y)| \leq \epsilon$$

Thus, (x^n, y^n) are jointly typical with respect to $Q(x,y)$ if the $P_n \in E$, 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 of a sequence being typical when it is generated from the product distribution Q_0 is

$$Q_0^n(P_n \in E) \approx 2^{-nD(P^*||Q_0)}$$

where P^* is the distribution satisfying the constraints that is closest to Q_0 in relative entropy. In this case, as $\epsilon \to 0$, we will verify 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 $2^{-nD(Q(x,y)||Q(x)Q(y))} = 2^{-nI(X;Y)}$.

To find $P^* = \arg\min_{P\in E} D(P||Q_0)$we use Lagrange multipliers and construct the Lagrangian function (for $\epsilon = 0$ and dropping the terms that don't depend on P)

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

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

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

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 $P^* = Q(x,y)$.

Thus, the probability that a sample of n data points drawn from Q_0 will appear jointly typical according to Q depends asymptotically scales as $2^{-nD(Q\|Q_0)}$. Similar statements can be made for general independence testing using Sanov's theorem.

The proof of Sanov's theorem for the finite-valued alphabet proceeds via Method of types which is a useful tool for many problems. We discuss this first.

23.2 Method of types

The method of types lets us associate a type (empirical distribution) to each sequence, and then by evaluating how many sequences have a particular type (empirical distribution), we can bound the probability of events. We start with a few definitions.

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

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

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 $T(P)$:

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

23.3 Some results using method of types

We establish some results using the method of types that will be useful for proving Sanov's theorem and also provide insights into the power of method of types. The first results just bounds the cardinality of set of types.

Theorem 23.2

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

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 a sufficient good upper bound. ■

The second result tells us that the probability of a sequence depends only on its type.

Theorem 23.3 *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:

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

■

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

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

The third result bounds the number of sequences of a given type.

Theorem 23.4 *(Size of a type class T(P)) For any type $P \in \mathcal{P}_n$,*

$$\frac{1}{(n+1)^{|\chi|}} 2^{nH(P)} \le |T(P)| \le 2^{nH(P)}$$

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

The upper bound follows by considering $P^n(T(P)) \le 1$ and bounding this by the size of the type class and 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.

The final result characterizes the probability of all sequences of a given type.

Theorem 23.5 *(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,*

$$\frac{1}{(n+1)^{|\chi|}} 2^{-nD(P||Q)} \le Q^n(T(P)) \le 2^{-nD(P||Q)}$$

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 $|T(P)|$, we have the following results

$$\frac{1}{(n+1)^{|\chi|}} 2^{-nD(P||Q)} \le Q^n(T(P)) \le 2^{-nD(P||Q)}$$

■

Equipped with these tools, we can now prove Sanov's theorem.

23.4 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:

$$\begin{aligned} Q^n(P_n \in E) &= \sum_{P \in E} Q^n((X_1, \ldots, X_n) \in T(P)) = \sum_{P \in E \cap \mathcal{P}_n} Q^n(T(P)) \\ &\leq \sum_{P \in E \cap \mathcal{P}_n} 2^{-nD(P||Q)} \leq |E \cap \mathcal{P}_n| \max_{P \in E \cap \mathcal{P}_n} 2^{-nD(P||Q)} \\ &\leq |\mathcal{P}_n| \; 2^{-n \min_{P \in E} D(P||Q)} \leq (n+1)^{|\mathcal{X}|} \; 2^{-nD(P^*||Q)} \end{aligned}$$

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

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

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

$$Q^n(P_n \in E) \geq Q^n(T(P'_n)) \geq \frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-nD(P'_n||Q)}$$

This implies that

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

which completes the proof.

23.5 Hypothesis Testing

Lets next apply the Sanov's theorem to bound the error probability in 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 $\min_g \beta_n P(H_1) + \alpha_n P(H_0)$.

The following theorem (stated without proof) characterizes the optimal test under Neyman-Pearson setting, which is essentially the likelihood ratio test.

Theorem 23.6 (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 $\beta >= \beta^*$.

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

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

23.5.1 Information-theoretic interpretation

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

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

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 23.1.

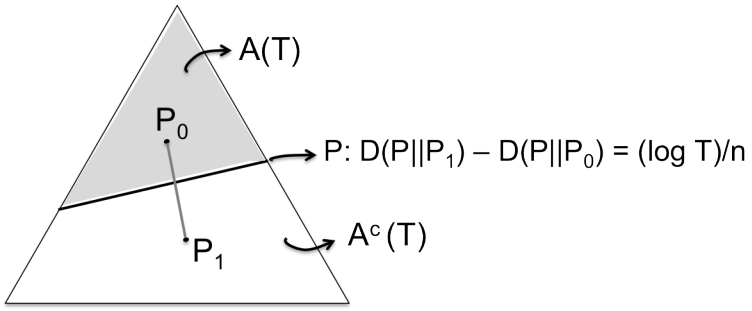

Figure 23.1: The decision region corresponding to a likelihood ratio test is demarcated by boundary that is parallel to the perpendicular bisector of the line joining the distributions under the two hypotheses P_0 and P_1.

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).

23.5.2 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^*).