

## Lecture 8: January 30, 2025

Lecturer: Madhur Tulsiani

## 1 Sanov's theorem

In the previous lecture, we derived Chernoff-Hoeffding bounds with the "right" exponent in terms of KL-divergence. This can actually be generalized to understand the probability that  $P_{\bar{x}} \in \Pi$  for an arbitrary set  $\Pi$  of distributions over  $\mathcal{X}$ . The following theorem gives the right exponent for all such concentration inequalities!

**Theorem 1.1** (Sanov). *Let  $\Pi$  be a set of distributions on  $\mathcal{X}$ , and  $|\mathcal{X}| = r$ . Then*

$$\mathbb{P}_{Q^n} [P_{\bar{x}} \in \Pi] \leq (n+1)^r \cdot 2^{-n \cdot \delta},$$

where  $\delta = \inf_{P \in \Pi} D(P \| Q)$ . Moreover, if  $\Pi$  is the closure of an open set and

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

then

$$\frac{1}{n} \cdot \log \left( \mathbb{P}_{\bar{x} \sim Q^n} [P_{\bar{x}} \in \Pi] \right) \rightarrow -D(P^* \| Q).$$

**Proof:** For any  $P \in \mathcal{T}_n$ , we have by Theorem ?? that

$$\mathbb{P}_{Q^n} [\bar{x} \in \mathcal{C}_P] \leq 2^{-nD(P \| Q)}.$$

Let  $\mathcal{T}_\delta = \{P \in \mathcal{T}_n \mid D(P \| Q) \geq \delta\}$ . Then, we have

$$\mathbb{P}_{\bar{x} \sim Q^n} [D(P_{\bar{x}} \| Q) \geq \delta] = \sum_{P \in \mathcal{T}_\delta} 2^{-n \cdot D(P \| Q)} \leq (n+1)^r \cdot 2^{-n\delta}.$$

We now use this to prove Sanov's theorem. Take  $\delta = \inf_{P \in \Pi} D(P \| Q)$ , so for all  $P \in \Pi$  we have  $D(P \| Q) \geq \delta$ . Then we get

$$\mathbb{P}_{\bar{x} \sim Q^n} [P_{\bar{x}} \in \Pi] = \mathbb{P}_{Q^n} [P_{\bar{x}} \in \Pi \cap \mathcal{T}_n] \leq \mathbb{P}_{Q^n} [D(P_{\bar{x}} \| Q) \geq \delta] \leq (n+1)^r \cdot 2^{-n\delta}$$

as desired. Now let's prove the other direction. Since  $\Pi$  is the closure of an open set (obtained by including the limit points of all converging sequences), we can say that the limit of the sequence converging to  $\inf_{P \in \Pi} D(P \| Q)$  exists in the set, and there exists  $P^* \in \Pi$  such that  $D(P^* \| Q) = \inf_{P \in \Pi} D(P \| Q)$ . This is the distribution  $P^*$  satisfying  $P^* := \arg \min_{P \in \Pi} D(P \| Q)$ .

Also, there is an  $n_0$  such that we can find a sequence  $\{P^{(n)}\}_{n \geq n_0}$  satisfying  $P^{(n)} \rightarrow P^*$  and  $P^{(n)} \in \mathcal{T}_n \cap \Pi$  for each  $n$ . Then we have

$$\begin{aligned} \mathbb{P}_{\bar{x} \sim Q^n} [P_{\bar{x}} \in \Pi] &= \mathbb{P}_{\bar{x} \sim Q^n} [P_{\bar{x}} \in \Pi] = \mathbb{P}_{\bar{x} \sim Q^n} [P_{\bar{x}} \in \Pi \cap \mathcal{T}_n] \\ &\geq \mathbb{P}_{\bar{x} \sim Q^n} [P_{\bar{x}} = P^{(n)}] \\ &\geq \frac{1}{(n+1)^r} \cdot 2^{-nD(P^{(n)} \| Q)} \end{aligned}$$

Thus we get

$$-D(P^{(n)} \| Q) - \frac{r \log(n+1)}{n} \leq \frac{1}{n} \log \left( \mathbb{P}_{\bar{x} \sim Q^n} [P_{\bar{x}} \in \Pi] \right) \leq -D(P^* \| Q) + \frac{r \log(n+1)}{n}$$

which gives

$$\frac{1}{n} \cdot \mathbb{P}_{Q^n} [P_{\bar{x}} \in \Pi] \rightarrow -D(P^* \| Q),$$

as claimed. ■

Note that the upper bound on the probability in Sanov's theorem holds for any  $\Pi$ . However, for the lower bound we need some conditions on  $\Pi$ . This is necessary since if (for example)  $\Pi$  is a set of distributions such that all probabilities in all the distributions are irrational, then  $\mathbb{P}_{Q^n} [P_{\bar{x}} \in \Pi] = 0$ . In particular, we cannot get any lower bound on this probability for such a  $\Pi$ .

Sanov's theorem can also be extended to the case when  $\mathcal{X}$  is an infinite set, using the definition of KL-divergence as a supremum over all finite partitions of an infinite space.

## 2 Binary hypothesis testing

We will now apply the tools developed in the past few lectures to understand the problem of distinguishing two distributions (special cases of which have been discussed in the previous lectures). This problem is also known as the hypothesis testing. Suppose we have two distributions  $P_0$  and  $P_1$  on a finite set  $\mathcal{X}$ . The "universe" chooses one of the two distributions and generates the data, which consists of a sequence  $\bar{x} \in \mathcal{X}^n$  chosen either from

$P_0^n$  or  $P_1^n$ . The true distribution is unknown to us, but we are guaranteed that once  $P_0$  or  $P_1$  is chosen, all  $n$  samples in the sequence  $\bar{x}$  are sampled independently from the chosen distribution. The goal is to distinguish between the following two hypotheses:

- $\mathbf{H}_0$ : The true distribution is  $P_0$ .
- $\mathbf{H}_1$ : The true distribution is  $P_1$ .

Sometimes  $\mathbf{H}_0$  is also referred to as the null (default) hypothesis. We will consider (deterministic) tests  $T : \mathcal{X}^n \rightarrow \{0, 1\}$ , which take the sequence of samples  $\bar{x}$  as input and select one of the hypotheses. There are two types of errors we will be concerned with

$$\begin{aligned}\alpha(T) &:= \mathbb{P}_{\bar{x} \sim P_0^n} [T(\bar{x}) = 1] \quad (\text{False Positive}) \\ \beta(T) &:= \mathbb{P}_{\bar{x} \sim P_1^n} [T(\bar{x}) = 0] \quad (\text{False Negative}).\end{aligned}$$

The following claim is easy to prove based on the properties of total-variation distance considered earlier.

**Proposition 2.1.**  $\min_T \{\alpha(T) + \beta(T)\} = 1 - \delta_{TV}(P_0^n, P_1^n) = 1 - \frac{1}{2} \cdot \|P_0^n - P_1^n\|_1.$

**Proof:** For any test  $T$ , we have,

$$\alpha(T) = \mathbb{E}_{\bar{x} \sim P_0^n} [T(\bar{x})] \quad \text{and} \quad \beta(T) = 1 - \mathbb{E}_{\bar{x} \sim P_1^n} [T(\bar{x})]$$

Thus,

$$\alpha(T) + \beta(T) = 1 - \left( \mathbb{E}_{\bar{x} \sim P_1^n} [T(\bar{x})] - \mathbb{E}_{\bar{x} \sim P_0^n} [T(\bar{x})] \right) \geq 1 - \delta_{TV}(P_0^n, P_1^n).$$

Also, recall that a test of the form

$$T(\bar{x}) = \begin{cases} 1 & \text{if } P_1^n(\bar{x}) \geq P_0^n(\bar{x}) \\ 0 & \text{if } P_1^n(\bar{x}) < P_0^n(\bar{x}) \end{cases},$$

is tight for the above inequality. ■

One may ask why should we only consider the optimal tests for minimizing the sum  $\alpha(T) + \beta(T)$ . We may care more about a false positive than a false negative, and may want to minimize a weighted sum (or some other monotone function) of the errors. Moreover, while the bound in [Proposition 2.1](#) (often computed using Pinsker's inequality) is useful in the case when  $\alpha(T)$  and  $\beta(T)$  are constants, it is harder to use when  $n$  is large and  $\alpha(T), \beta(T)$  are decreasing (exponentially) with  $n$ .

We will in fact be able to characterize an optimal family of tests, and obtain bounds on  $\alpha(T)$  and  $\beta(T)$  individually. The following lemma shows that all optimal tests should be of the form above, which make a decision only based on the *ratio*  $P_0^n(\bar{x})/P_1^n(\bar{x})$ .

**Lemma 2.2** (Neyman-Pearson Lemma). *Let  $T$  be a test of the form*

$$T(\bar{\mathbf{x}}) = \begin{cases} 1 & \text{if } P_1^n(\bar{\mathbf{x}})/P_0^n(\bar{\mathbf{x}}) \geq \Delta \\ 0 & \text{if } P_0^n(\bar{\mathbf{x}})/P_1^n(\bar{\mathbf{x}}) < \Delta, \end{cases}$$

for some constant  $\Delta > 0$ . Let  $T'$  be any other test. Then,

$$\alpha(T') \geq \alpha(T) \quad \text{or} \quad \beta(T') \geq \beta(T).$$

**Proof:** The proof follows simply from the observation that for all  $\bar{\mathbf{x}} \in \mathcal{X}^n$

$$(T(\bar{\mathbf{x}}) - T'(\bar{\mathbf{x}})) \cdot (P_1^n(\bar{\mathbf{x}}) - \Delta \cdot P_0^n(\bar{\mathbf{x}})) \geq 0.$$

This is true because if  $P_1^n(\bar{\mathbf{x}}) - \Delta \cdot P_0^n(\bar{\mathbf{x}}) \geq 0$ , then  $T(\bar{\mathbf{x}}) = 1$  and the first quantity is non-negative. Similarly, when  $P_1^n(\bar{\mathbf{x}}) - \Delta \cdot P_0^n(\bar{\mathbf{x}})$  is negative,  $T(\bar{\mathbf{x}}) = 0$  and  $T(\bar{\mathbf{x}}) - T'(\bar{\mathbf{x}}) \leq 0$ . Summing over all  $\bar{\mathbf{x}} \in \mathcal{X}^n$  on both sides gives

$$\begin{aligned} & \mathbb{E}_{\bar{\mathbf{x}} \sim P_1^n} [T(\bar{\mathbf{x}}) - T'(\bar{\mathbf{x}})] - \Delta \cdot \mathbb{E}_{\bar{\mathbf{x}} \sim P_0^n} [T(\bar{\mathbf{x}}) - T'(\bar{\mathbf{x}})] \geq 0 \\ \Rightarrow & ((1 - \beta(T)) - (1 - \beta(T'))) - \Delta \cdot (\alpha(T) - \alpha(T')) \geq 0 \\ \Rightarrow & \frac{\beta(T') - \beta(T)}{\alpha(T) - \alpha(T')} \geq \Delta > 0. \end{aligned}$$

Thus,  $\alpha(T) - \alpha(T') \geq 0$  implies  $\beta(T') - \beta(T) \geq 0$ . ■

## 2.1 Analyzing error exponents

We now discuss how to analyze the error probabilities for the optimal tests as characterized by the Neyman-Pearson lemma. As before, let  $P_{\bar{\mathbf{x}}}$  denote the type (empirical distribution on  $\mathcal{X}$ ) of the sequence  $\bar{\mathbf{x}}$ . Check that the test  $T(\bar{\mathbf{x}})$  considered above can be written in the following form

$$\frac{P_1^n(\bar{\mathbf{x}})}{P_0^n(\bar{\mathbf{x}})} \geq \Delta \quad \Leftrightarrow \quad D(P_{\bar{\mathbf{x}}} \| P_0) - D(P_{\bar{\mathbf{x}}} \| P_1) \geq \frac{1}{n} \cdot \log \Delta.$$

We define the following sets of probability distributions.

$$\begin{aligned} \Pi & := \left\{ P \mid D(P \| P_0) - D(P \| P_1) \geq \frac{1}{n} \cdot \log \Delta \right\} \\ \Pi^c & := \left\{ P \mid D(P \| P_0) - D(P \| P_1) < \frac{1}{n} \cdot \log \Delta \right\} \end{aligned}$$

Check the following property of the sets  $\Pi$  and  $\Pi^c$ .

**Exercise 2.3.** Check that both the sets  $\Pi$  and  $\Pi^c$  are convex (and are in fact defined by linear inequalities in the distributions  $P$ ). Also, check that  $\Pi$  is a closed set.

We can now estimate  $\alpha(T)$  and  $\beta(T)$  using Sanov's theorem. The  $\approx$  notation below ignores second order terms in the exponents. We get

$$\begin{aligned}\alpha(T) &= \mathbb{P}_{\bar{x} \sim P_0^n} [P_{\bar{x}} \in \Pi] \approx 2^{-n \cdot D(P_0^* \| P_0)} \\ \beta(T) &= \mathbb{P}_{\bar{x} \sim P_1^n} [P_{\bar{x}} \in \Pi^c] \approx 2^{-n \cdot D(P_1^* \| P_1)},\end{aligned}$$

where  $P_0^* = \arg \min_{P \in \Pi} \{D(P \| P_0)\}$ . Also, since  $\Pi^c$  is not a closed set, we define  $P_1^*$  with respect to the closure of  $\Pi^c$  of  $\Pi^c$  i.e.,  $P_1^* = \arg \min_{P \in \overline{\Pi^c}} \{D(P \| P_1)\}$ .

We will see in a later lecture how to compute the distributions which minimize the KL-divergence (known as I-projections) as in the bounds above. The distributions  $P_0^*$  and  $P_1^*$  in the above bounds turn out to be the same, and of the form

$$P_0^*(x) = P_1^*(x) = P^* = \frac{P_0^\lambda(x) \cdot P_1^{1-\lambda}(x)}{\sum_{y \in \mathcal{X}} P_0^\lambda(y) \cdot P_1^{1-\lambda}(y)},$$

where  $\lambda$  is the solution to an optimization problem.

Note that since  $P_1 \in \Pi$  ( $T$  will always answer 0), we have that  $D(P_0^* \| P_0) \leq D(P_1 \| P_0)$ . Also, as  $\Delta$  increases, the boundary of  $\Pi$  approaches closer to  $P_1$  and the exponent approaches  $D(P_1 \| P_0)$ . This is made precise by the Chernoff-Stein lemma, which we will not discuss in detail (but is good to know).

## 2.2 Bayesian error

Going back to the case of the expression  $\alpha(T) + \beta(T)$ , we can view it as (twice) the expected error in case we have a prior distribution on the hypotheses. If we chose the hypotheses with probabilities 1/2 each, the expected error will be  $\frac{1}{2} \cdot (\alpha(T) + \beta(T))$ . Recall that the optimal test in this case used  $\Delta = 1$  i.e.,

$$T(\bar{x}) = \begin{cases} 1 & \text{if } P_1^n(\bar{x})/P_0^n(\bar{x}) \geq 1 \\ 0 & \text{if } P_0^n(\bar{x})/P_1^n(\bar{x}) < 1 \end{cases} = \arg \min_{i \in \{0,1\}} \{D(P_{\bar{x}} \| P_i)\},$$

and the sets  $\Pi$  and  $\Pi^c$  are defined as

$$\begin{aligned}\Pi &:= \{P \mid D(P \| P_0) - D(P \| P_1) \geq 0\} \\ \Pi^c &:= \{P \mid D(P \| P_0) - D(P \| P_1) < 0\}\end{aligned}$$

When  $n$  is large enough, we will have

$$\frac{1}{2} \cdot (\alpha(T) + \beta(T)) \approx 2^{-n \cdot D(P^* \| P_0)} + 2^{-n \cdot D(P^* \| P_1)} \approx 2^{-n \cdot \min\{D(P^* \| P_0), D(P^* \| P_1)\}},$$

where  $P^* = \arg \min_{P \in \Pi} D(P \| P_0) = \arg \min_{P \in \overline{\Pi}^c} D(P \| P_1)$ , as before. Also note that the exponent remains the same when considering *any* prior distribution  $(\pi, 1 - \pi)$  (not dependent on  $n$ ) on the hypotheses  $P_0$  and  $P_1$ , as the first order term in the exponent is still proportional to  $C(P_0, P_1) := \min\{D(P^* \| P_0), D(P^* \| P_1)\}$ . This quantity, which is symmetric in terms of  $P_0$  and  $P_1$  is referred to as the Chernoff distance between the two distributions, and is said to define the optimal exponent for the Bayesian error of hypothesis testing.

Note that the above analysis is only for the case of large  $n$ . When  $n$  is small, the bound we will use the most is the lower bound in terms of the total variation distance i.e.,

$$\min_T \{\alpha(T) + \beta(T)\} = 1 - \delta_{TV}(P_0^n, P_1^n).$$