Information and Coding Theory **Winter 2021**

Lecture 8: February 4, 2021

Lecturer: Madhur Tulsiani

1 Binary hypothesis testing

In this lecture, we 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 $\overline{\mathbf{x}} \in \mathcal{X}^n$ 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 $\overline{\mathbf{x}}$ 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 $T : \mathcal{X}^n \to \{0, 1\}$, which take the sequence of samples $\overline{\mathbf{x}}$ as input and select one of the hypotheses. There are two types of errors we will be concerned with

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

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

Proposition 1.1. $\min_T \{\alpha(T) + \beta(T)\} \;=\; 1 - \delta_{TV}\left(P_0^n, P_1^n\right) \;=\; 1 - \frac{1}{2} \cdot \|P_0^n - P_1^n\|_1$.

Proof: For any test T, we have,

$$\alpha(T) \;=\; \underset{\overline{\mathbf{x}} \sim P_0^n}{\mathbb{E}} \left[T(\overline{\mathbf{x}}) \right] \quad \text{and} \quad \beta(T) \;=\; 1 - \underset{\overline{\mathbf{x}} \sim P_1^n}{\mathbb{E}} \left[T(\overline{\mathbf{x}}) \right]$$

Thus,

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

Also, recall that a test of the form

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

is tight for the above inequality. ■

One may ask why should be 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 1.1 (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* $P_0^n(\overline{\mathbf{x}})/P_1^n(\overline{\mathbf{x}})$.

Lemma 1.2 (Neyman-Pearson Lemma)**.** *Let T be a test of the form*

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

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

$$\alpha(T') \;\geq\; \alpha(T) \qquad or \qquad \beta(T') \;\geq\; \beta(T) \,.$$

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

$$\left(T(\overline{\mathbf{x}}) - T'(\overline{\mathbf{x}})\right) \cdot \left(P_1^n(\overline{\mathbf{x}}) - \Delta \cdot P_0^n(\overline{\mathbf{x}})\right) \;\geq\; 0 \,.$$

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

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

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

1.1 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 $P_{\overline{\mathbf{x}}}$ denote the type (empirical distribution on $\mathcal{X}$) of the sequence $\overline{\mathbf{x}}$. Check that the test $T(\overline{\mathbf{x}})$ considered above can be written in the following form

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

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 Π and Π^c.

Exercise 1.3. *Check that both the sets Π and Π^c are convex (and are in fact defined by linear inequalities in the distributions P). Also, check that Π 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

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

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

1.2 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.,

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

and the sets Π and Π^c are defined as

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

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.,

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

We will also develop such a bound for the case of multiple hypotheses.

2 Multiple hypothesis testing

We will often use the case of teting between multiple hypotheses as proof technique for lower bounds, and the important bound there will be an analog of the bound for small n in case of the binary hypothesis testing. However, before that we briefly discuss known generalizations of the results for binary hypotheses in the case of large n.

2.1 Bayesian error

Consider the case of distinguishing between k distributions $P_1, \ldots, P_k$ on $\mathcal{X}$, again using a sequence $\overline{\mathbf{x}} = (x_1, \ldots, x_n)$ of n independent samples from one of them. A test $T(\overline{\mathbf{x}})$ now needs to have an output in $[k]$ and can have $k(k-1)$ types of errors, of the form

$$\alpha_{ij} \;:=\; \underset{\overline{\mathbf{x}} \sim P_i^n}{\mathbb{P}} \left[T(\overline{\mathbf{x}}) = j\right] \,.$$

While it is harder to characterize the optimal error tests for each individual error type, a generalization of the Bayesian error analysis was obtained by Leang and Johnson [LJ97] (see also [Wes08] for a different interpretation of the test). Given any prior $(\pi_1, \ldots, \pi_k)$ on the k hypotheses, the Bayesian error is a sum of $k(k-1)$ terms, and is equal to

$$\pi_1 \cdot \left(\sum_{j \neq 1} \alpha_{1j}\right) + \cdots + \pi_k \cdot \left(\sum_{j \neq k} \alpha_{kj}\right)$$

As n increases, the exponential decay of the largest term among these dominates the error rate, and the exponent is proportional to $\min_{i \neq j} C(i,j)$, where $C(i,j)$ (Chernoff distance) is the optimal exponent for the binary case discussed above i.e., the error is dominated by the two hypotheses closest in the Chernoff distance. The optimal test for the Bayesian error is also a generalization of the binary case, and is of the form

$$T(\overline{\mathbf{x}}) \;=\; \underset{i \in [k]}{\arg\min} \left\{D(P_{\overline{\mathbf{x}}} \| P_i)\right\} \,.$$

We will not discuss (or need) the details of this case, but please see the references [LJ97, Wes08] for a proof.

2.2 Fano's inequality and a lower bound

We will now prove a lower bound on the error analogous to Proposition 1.1 in the binary case. This will rely on Fano's inequality, for which we recall the statement below.

Lemma 2.1 (Fano's inequality)**.** *Let $Z \to Y \to \widehat{Z}$ be a Markov chain with Z taking values in a finite set $\mathcal{Z}$, and let $p_e = \mathbb{P}\left[\widehat{Z} \neq Z\right]$. Let $H_2(p_e)$ denote the binary entropy function computed at p_e. Then,*

$$H_2(p_e) + p_e \cdot \log\left(|\mathcal{Z}| - 1\right) \;\geq\; H(Z|\widehat{Z}) \;\geq\; H(Z|Y) \,.$$

Let $\{P_v\}_{v \in \mathcal{V}}$ be a collection of hypotheses. Let the environment choose one of the hypotheses uniformly at random, denoted by a random variable V distributed uniformly in $\mathcal{V}$. Let $\overline{\mathbf{x}} \sim P_v^n$ be a sequence of independent samples from a chosen distribution P_v (denoted by the random variable $\overline{\mathbf{X}}$). We will now bound the probability of error for a classifier $\widehat{V}$ for V. Note that $V \to \overline{\mathbf{X}} \to \widehat{V}$ is a Markov chain.

Proposition 2.2. *Let $V \to \overline{\mathbf{X}} \to \widehat{V}$ be the Markov chain as above. Then,*

$$p_e \;=\; \mathbb{P}\left[V \neq \widehat{V}\right] \;\geq\; 1 - \frac{n \cdot \mathbb{E}_{v_1,v_2\in\mathcal{V}}\left[D(P_{v_1}\|P_{v_2})\right] + 1}{\log|\mathcal{V}|}\,.$$

Proof: From Fano's inequality, we have that

$$1 + p_e \cdot \log|\mathcal{V}| \;\geq\; H(p_e) + p_e \cdot \log|\mathcal{V}| \;\geq\; H(V|\overline{\mathbf{X}}) \;=\; \log|\mathcal{V}| - I(V;\overline{\mathbf{X}})\,.$$

We can now analyze the mutual information between V and $\bar{\mathbf{x}}$ using the equivalent expression in terms of KL-divergence.

$$\begin{aligned} I(V;\bar{\mathbf{x}}) &= D(P(V,\overline{\mathbf{X}})\|P(V)P(\overline{\mathbf{X}})) \\ &= D(P(V)\|P(V)) + \mathop{\mathbb{E}}_{v\in\mathcal{V}}\left[D(P(\overline{\mathbf{X}}|V=v)\|P(\overline{\mathbf{X}}))\right] \\ &= \mathop{\mathbb{E}}_{v\in\mathcal{V}}\left[D(P_v^n\|\overline{P})\right]\,, \end{aligned}$$

where $\overline{P} = \mathbb{E}_{v\in\mathcal{V}}\left[P_v^n\right]$ denotes the marginal distribution of $\overline{\mathbf{X}}$. Using the convexity of KL-divergence in the second argument and Jensen's inequality, we get

$$\mathop{\mathbb{E}}_{v\in\mathcal{V}}\left[D(P_v^n\|\overline{P})\right] \;\leq\; \mathop{\mathbb{E}}_{v_1,v_2\in\mathcal{V}}\left[D(P_{v_1}^n\|P_{v_2}^n)\right]\,.$$

Using the chain rule for KL-divergence gives

$$\mathop{\mathbb{E}}_{v_1,v_2\in\mathcal{V}}\left[D(P_{v_1}^n\|P_{v_2}^n)\right] \;=\; n \cdot \mathop{\mathbb{E}}_{v_1,v_2\in\mathcal{V}}\left[D(P_{v_1}\|P_{v_2})\right]\,.$$

Combining the bounds, we have

$$1 + p_e \cdot \log|\mathcal{V}| \;\geq\; \log|\mathcal{V}| - n \cdot \mathop{\mathbb{E}}_{v_1,v_2\in\mathcal{V}}\left[D(P_{v_1}\|P_{v_2})\right]\,,$$

which proves the claim. ■

References

[LJ97] Charles C Leang and Don H Johnson, *On the asymptotics of M-hypothesis Bayesian detection*, IEEE Transactions on Information Theory **43** (1997), no. 1, 280–282. 5

[Wes08] M Brandon Westover, *Asymptotic geometry of multiple hypothesis testing*, IEEE transactions on information theory **54** (2008), no. 7, 3327–3329. 5