Information and Coding Theory **Autumn 2017**

Lecture 7: October 18, 2017

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 universe U. The "universe" chooses one of the two distributions and generates the dats, which consists of a sequence $\overline{\mathbf{x}} \in U^n$ 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 $T : U^n \to \{0,1\}$, 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

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

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

Claim 1.1 $\min_T \{\alpha(T) + \beta(T)\} \;=\; 1 - \delta_{TV}\left(P_0^n, P_1^n\right)$.

Recall that optimal test for the above claim should be of the form

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

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. 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 $\Delta \geq 0$. Let T' be any other test. Then,

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

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

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

This is true because if $P_1^n(\overline{\mathbf{x}}) - \Delta \cdot P_0^n(\overline{\mathbf{x}})$, 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 $\overline{\mathbf{x}} \in U^n$ on both sides gives

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

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

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 U) 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 know from Sanov's theorem that

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

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 later 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 of the form

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

where λ is the solution to an optimization problem. While the above analysis gives the optimal bounds for optimal all tests characterized by the Neyman-Pearson lemma, the bound we will use the most is the lower bound in terms of the total variation distance i.e.,

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

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

2 Fano's inequality and multiple hypothesis testing

Fano's inequality is concerned with Markov chains, which we saw before in the context of data processing inequality. We will denote the Markov chain as $Z \to Y \to \widehat{Z}$. In the context of hypothesis testing, we can think of Z as the choice of an unknown hypothesis from some finite set (hypothesis class) U_Z. We think of Y as the "data" generated from this hypothesis, say a sequence $\overline{\mathbf{x}}$ of n independent samples. Finally, we think of $\widehat{Z}$ as a "guess" for Z, which depends only on the data. Fano's inequality is concerned with the probability of error in the guess, defined as $p_e = \mathbb{P}\left[\widehat{Z} \neq Z\right]$. We have the following statement

Lemma 2.1 (Fano's inequaity) *Let $Z \to Y \to \widehat{Z}$ be a Markov chain, and let $p_e = \mathbb{P}\left[\hat{Z} \neq Z\right]$. Let $H(p_e)$ denote the binary entropy function computed at p_e. Then,*

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

Proof: We define a binary random variable, which indicates an error i.e

$$E \;:=\; \begin{cases} 1 \text{ if } \widehat{Z} \neq Z \\ 0 \text{ if } \widehat{Z} = Z \end{cases}$$

The bound in the ineuality then follows from considering the entroy $H(Z, E|\widehat{Z})$.

$$H(Z, E|\widehat{Z}) \;=\; H(Z|\widehat{Z}) + H(E|Z,,\widehat{Z}) \;=\; H(Z|\widehat{Z})\,,$$

since $H(E|Z, \widehat{Z}) = 0$ (why?) Another way of computing this entropy is

$$\begin{aligned} H(Z, E|\widehat{Z}) \;&=\; H(E|\widehat{Z}) + H(Z|E, \widehat{Z}) \\ &=\; H(E|\widehat{Z}) + p_e \cdot H(Z|E = 1, \widehat{Z}) + (1 - p_e) \cdot H(Z|E = 0, \widehat{Z}) \\ &\leq\; H(E) + p_e \cdot H(Z|E = 1, \widehat{Z}) \\ &\leq\; H(p_e) + p_e \cdot \log\left(|U_Z| - 1\right)\,. \end{aligned}$$

Comparing the two expressions them proves the claim. ∎

We can use Fano's inequality to derive a convenient way of obtaining a lower bound for testing multiple hypotheses. However, we need the following property of KL-divergence.

Exercise 2.2 *Prove that KL-divergence is (strictly) convex in both it's arguments i.e., $\forall \alpha \in (0, 1)$ and all $P_1 \neq P_2$, $Q_1 \neq Q_2$,*

$$\begin{aligned} D(\alpha \cdot P_1 + (1 - \alpha) \cdot P_2 \| Q) \;&<\; \alpha \cdot D(P_1 \| Q) + (1 - \alpha) \cdot D(P_2 \| Q) \\ D(P \| \alpha \cdot Q_1 + (1 - \alpha) \cdot Q_2) \;&<\; \alpha \cdot D(P \| Q_1) + (1 - \alpha) \cdot D(P \| Q_2) \end{aligned}$$

In fact, KL-divergence is jointly *convex in both its arguments but we will need this property.*

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 vaiable V) and let $\overline{\mathbf{x}} \sim P_v^n$ be a sequence of independent samples from a chosen disribution 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 2.3 *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 $\overline{\mathbf{x}}$ using the equivalent expression in terms of KL-divergence.

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

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

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

Combining the bounds gives

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

which proves the claim. ∎