

Lecture 9: February 6, 2025

Lecturer: Madhur Tulsiani

In this lecture, we will use lower bounds on hypothesis testing developed before to understand how well we can “learn” properties of distributions using samples. Much of the presentation here is based on the excellent set of lecture notes by John Duchi [Duc16] (also linked from the course webpage) which I highly recommend for a more in-depth treatment of the subject.

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

1 Multiple hypothesis testing

We will often use the case of testing 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 .

1.1 Bayesian error

We did not discuss the following in class, but just adding some notes here in case you want to compare this to the results known for binary hypothesis testing. Consider the case of distinguishing between k distributions P_1, \dots, P_k on \mathcal{X} , again using a sequence $\bar{x} = (x_1, \dots, x_n)$ of n independent samples from one of them. A test $T(\bar{x})$ now needs to have an output in $[k]$ and can have $k(k-1)$ types of errors, of the form

$$\alpha_{ij} := \mathbb{P}_{\bar{x} \sim P_i^n} [T(\bar{x}) = j].$$

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 (π_1, \dots, π_k) on the k hypotheses, the Bayesian error is a sum of $k(k-1)$ terms, and is equal to

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

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

$$T(\bar{\mathbf{x}}) = \arg \min_{i \in [k]} \{D(P_{\bar{\mathbf{x}}} \| P_i)\}.$$

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

1.2 Fano's inequality and a lower bound

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

Lemma 1.1 (Fano's inequality). *Let $Z \rightarrow Y \rightarrow \hat{Z}$ be a Markov chain with Z taking values in a finite set \mathcal{Z} , and let $p_e = \mathbb{P}[\hat{Z} \neq Z]$. Let $H_2(p_e)$ denote the binary entropy function computed at p_e . Then,*

$$H_2(p_e) + p_e \cdot \log(|\mathcal{Z}| - 1) \geq H(Z|\hat{Z}) \geq H(Z|Y).$$

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 $\bar{\mathbf{x}} \sim P_v^n$ be a sequence of independent samples from a chosen distribution P_v (denoted by the random variable $\bar{\mathbf{X}}$). We will now bound the probability of error for a classifier \hat{V} for V . Note that $V \rightarrow \bar{\mathbf{X}} \rightarrow \hat{V}$ is a Markov chain.

Proposition 1.2. *Let $V \rightarrow \bar{\mathbf{X}} \rightarrow \hat{V}$ be the Markov chain as above. Then,*

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

Proof: From Fano's inequality, we have that

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

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

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

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

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

Using the chain rule for KL-divergence gives

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

Combining the bounds, we have

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

which proves the claim. ■

2 Minimax risk and reduction to hypothesis testing

Let Π be a set of distributions on U and let $\theta : \Pi \rightarrow \Theta$ be any map which we think as a "property" of P . We consider an estimator $\hat{\theta} : \mathcal{X}^n \rightarrow \Theta$, which takes n independent samples from P as input, and tries to estimate $\theta(P)$. The quality of the estimator is measured by a *loss function* $\ell : \Theta \times \Theta \rightarrow \mathbb{R}_+$. If we use an estimator $\hat{\theta}$ and the data comes from a distribution P , the *expected loss* is $\mathbb{E}_{\bar{\mathbf{x}} \sim P^n} [\ell(\hat{\theta}(\bar{\mathbf{x}}), \theta(P))]$. The goal is to come up with an estimator, which minimizes the loss even for the worst-case distribution i.e., we want to understand

$$\mathcal{M}_n(\Pi, \ell) := \inf_{\hat{\theta}} \sup_{P \in \Pi} \mathbb{E}_{\bar{\mathbf{x}} \sim P^n} [\ell(\hat{\theta}(\bar{\mathbf{x}}), \theta(P))]$$

The quantity $\mathcal{M}_n(\Pi, \ell)$ is also called the *minimax risk*. As an example, consider the case $\Pi = \{P_v\}_{v \in \mathcal{V}}$, $\Theta = \mathcal{V}$ and $\theta(P_v) = v$. We take $\ell(\hat{\theta}, \theta) = 1$ if $\hat{\theta} \neq \theta$ and 0 otherwise. The goal is to find

$$\mathcal{M}_n(\Pi, \ell) = \inf_{\hat{\theta}} \sup_{v \in \mathcal{V}} \mathbb{P}_{\bar{\mathbf{x}} \sim P_v^n} [\hat{\theta}(\bar{\mathbf{x}}) \neq v] ,$$

which is very similar to the setting of multiple hypothesis testing introduced in the previous lecture. While the minimax risk requires bounding the probability of error for the *worst* distribution in Π , in the previous lecture we developed a lower bound on the probability that the estimator errs for a *randomly chosen* distribution from Π . Of course this is still a lower bound. If we have some additional information about \mathcal{V} , we can find a "hard set" $\Pi' \subseteq \Pi$ and apply the bound from the previous lecture for a randomly chosen distribution from Π' . This is still a lower bound on the minimax risk. All the lower bounds

developed below are essentially of this form, where we identify a hard subset of distributions and apply the bounds for hypothesis testing. In general, the notion of a “hard subset” of distributions needs to be developed with respect to the loss function ℓ .

We will restrict the discussion here to loss functions ℓ which only depend on some form of distance between $\hat{\theta}$ and θ . In particular, we consider

$$\ell(\hat{\theta}, \theta) = \Phi(\rho(\hat{\theta}, \theta)) = \Phi \circ \rho(\hat{\theta}, \theta),$$

where $\rho(\cdot, \cdot)$ is a metric (obeying triangle inequality) and Φ is a non-negative and non-decreasing function. In fact, $\ell(\hat{\theta}, \theta) = \|\hat{\theta} - \theta\|_2^2$ will suffice for our purposes, but we state the reduction from lower bounds on minimax risk to hypothesis testing for any ℓ of the form above.

Lemma 2.1. *Let $\{P_v\}_{v \in \mathcal{V}} \subseteq \Pi$ be a finite set of distributions such that $\forall v_1, v_2 \in \mathcal{V}$ with $v_1 \neq v_2$, $\rho(\theta(P_{v_1}), \theta(P_{v_2})) \geq 2\delta$. Let ℓ be as above. Then,*

$$\mathcal{M}(\Pi, \ell) \geq \Phi(\delta) \cdot \inf_T \{\mathbb{P}[T(\bar{\mathbf{x}}) \neq V]\}.$$

Note that the setting in the RHS above is exactly as considered in hypothesis testing. We think of V as uniformly distributed over the set \mathcal{V} and $\bar{\mathbf{x}}$ as drawn from P_v^n .

Proof: Let $\hat{\theta} : U^n \rightarrow \mathcal{V}$ be any estimator. We define a classifier $T : U^n \rightarrow \mathcal{V}$ (depending on $\hat{\theta}$) as follows

$$T(\bar{\mathbf{x}}) := \arg \min_{v \in \mathcal{V}} d(\hat{\theta}(\bar{\mathbf{x}}), \theta(P_v)).$$

Note that if $V = v$ and $T(\bar{\mathbf{x}}) = v' \neq v$, we must have $d(\hat{\theta}(\bar{\mathbf{x}}), \theta(P_v)) \geq \delta$ (why?) This implies that if T makes an error on input $\bar{\mathbf{x}}$, then we must have $\ell(\hat{\theta}, \theta) \geq \Phi(\delta)$. Thus, we get

$$\begin{aligned} \sup_{P \in \Pi} \mathbb{E}_{\bar{\mathbf{x}} \sim P^n} [\Phi \circ \rho(\hat{\theta}(\bar{\mathbf{x}}), \theta(P))] &\geq \mathbb{E}_{v \in \mathcal{V}} \mathbb{E}_{\bar{\mathbf{x}} \sim P_v^n} [\Phi \circ \rho(\hat{\theta}(\bar{\mathbf{x}}), \theta(P_v))] \\ &\geq \Phi(\delta) \cdot \mathbb{P}[T(\bar{\mathbf{x}}) \neq V] \\ &\geq \Phi(\delta) \cdot \inf_T \{\mathbb{P}[T(\bar{\mathbf{x}}) \neq V]\}. \end{aligned}$$

The last inequality above used the fact that the error of the classifier T here is lower bounded by the error of the *best* classifier. Since after taking the infimum over T , the above bound now holds for *any* $\hat{\theta}$, it also we get that

$$\mathcal{M}_n(\Pi, \ell) = \inf_{\hat{\theta}} \sup_{P \in \Pi} \mathbb{E}_{\bar{\mathbf{x}} \sim P^n} [\Phi \circ \rho(\hat{\theta}(\bar{\mathbf{x}}), \theta(P))] \geq \Phi(\delta) \cdot \inf_T \{\mathbb{P}[T(\bar{\mathbf{x}}) \neq V]\},$$

which proves the claim. ■

3 Lower bounds via binary hypothesis testing (Le Cam's method)

We return to our favorite example of biased coins. Let $\mathcal{X} = \{0, 1\}$ and let Π be the set of all distributions on $\{0, 1\}$. For a distribution P on \mathcal{X} , let $\theta(P) := p(1) = \mathbb{E}_{x \sim P}[x]$ i.e., the goal is to estimate the probability that the coin comes up heads (the mean of a Bernoulli random variable). We first consider a very simple estimator, which just takes the empirical mean of the given data i.e.,

$$\hat{\theta}(\bar{x}) = \hat{\theta}(x_1, \dots, x_n) := \frac{1}{n} \cdot \sum_{i \in [n]} x_i.$$

Check that the expected error of this estimator, for the loss function $\ell(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$, is $O(1/n)$.

Exercise 3.1. Let $P : \{0, 1\} \rightarrow [0, 1]$ be any distribution with $\mathbb{E}_{x \sim P}[x] = p(1) = \mu$. Show that

$$\mathbb{E}_{(x_1, \dots, x_n) \sim P^n} \left[\left| \frac{1}{n} \cdot \sum_{i \in [n]} x_i - \mu \right|^2 \right] = O\left(\frac{1}{n}\right).$$

We will now show that the above bound is tight. Let $\mathcal{V} = \{0, 1\}$, and let $P_0 = (1/2, 1/2)$ and $P_1 = (1/2 - 2\delta, 1/2 + 2\delta)$ be the corresponding two distributions (the value of δ will be chosen later). Note that

$$|\theta(P_0) - \theta(P_1)| = 2\delta.$$

Using the lemma from the previous section, we get that

$$\begin{aligned} \mathcal{M}(\Pi, \ell) &\geq \delta^2 \cdot \inf_T \{ \mathbb{P}[T(\bar{x}) \neq V] \} \\ &\geq \delta^2 \cdot \inf_T \left\{ \frac{1}{2} \cdot \mathbb{P}_{\bar{x} \sim P_0^n} [T(\bar{x}) = 1] + \mathbb{P}_{\bar{x} \sim P_1^n} [T(\bar{x}) = 0] \right\} \\ &\geq \delta^2 \cdot \frac{1}{2} \cdot \inf_T \{ \alpha(T) + \beta(T) \}, \end{aligned}$$

where $\alpha(T)$ and $\beta(T)$ are the errors as defined in the setting of binary hypothesis testing. Using the bound in terms of total-variation distance, we get that

$$\begin{aligned} \mathcal{M}(\Pi, \ell) &\geq \frac{\delta^2}{2} \cdot \left(1 - \frac{1}{2} \cdot \|P_0^n - P_1^n\|_1 \right) \\ &\geq \frac{\delta^2}{2} \cdot \left(1 - \frac{1}{2} \cdot \sqrt{2 \ln 2 \cdot n \cdot D(P_0 \| P_1)} \right). \end{aligned}$$

We use the calculation from the previous lectures that $D(P_0 \| P_1) \leq c\delta^2$ for some constant c . Choosing $\delta = (c \cdot 2 \ln 2 \cdot n)^{-1/2}$ gives

$$\mathcal{M}(\Pi, \ell) \geq \frac{\delta^2}{2} \left(1 - \frac{1}{2} \right) = \Omega\left(\frac{1}{n}\right).$$

References

- [Duc16] John Duchi, *Lecture notes on Information Theory and Statistics*, 2016. [1](#)
- [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. [1](#), [2](#)
- [Wes08] M Brandon Westover, *Asymptotic geometry of multiple hypothesis testing*, *IEEE transactions on information theory* **54** (2008), no. 7, 3327–3329. [1](#), [2](#)