Information and Coding Theory **Autumn 2022**

Lecture 11 & 12: November 1 & 3, 2022

Lecturer: Omar Montasser

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.

1 Minimax risk and reduction to hypothesis testing

Let Π be a set of distributions on U and let $\theta : \Pi \to \Theta$ be any map which we think as a "property" of P. We consider an estimator $\widehat{\theta} : \mathcal{X}^n \to \Theta$, 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* $\ell : \Theta \times \Theta \to \mathbb{R}_+$. If we use an estimator $\widehat{\theta}$ and the data comes from a distribution P, the *expected* loss is $\mathbb{E}_{\overline{\mathbf{x}} \sim P^n}\left[\ell\left(\widehat{\theta}(\overline{\mathbf{x}}), \theta(P)\right)\right]$. 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

$$\mathcal{M}_n(\Pi, \ell) \;:=\; \inf_{\widehat{\theta}} \sup_{P \in \Pi} \mathop{\mathbb{E}}_{\overline{\mathbf{x}} \sim P^n}\left[\ell\left(\widehat{\theta}(\overline{\mathbf{x}}), \theta(P)\right)\right] .$$

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 $\ell(\widehat{\theta}, \theta) = 1$ if $\widehat{\theta} \neq \theta$ and 0 otherwise. The goal is to find

$$\mathcal{M}_n(\Pi, \ell) \;=\; \inf_{\widehat{\theta}} \sup_{v \in \mathcal{V}} \mathop{\mathbb{P}}_{\overline{\mathbf{x}} \sim P_v^n}\left[\theta(\overline{\mathbf{x}}) \neq v\right] ,$$

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 $\widehat{\theta}$ and θ. In particular, we consider

$$\ell\left(\widehat{\theta},\theta\right) \;=\; \Phi\left(\rho\left(\widehat{\theta},\theta\right)\right) \;=\; \Phi\circ\rho\left(\widehat{\theta},\theta\right) \;,$$

where $\rho(\cdot,\cdot)$ is a metric (obeying triangle inequality) and Φ is a non-negative and non-decreasing function. In fact, $\ell(\widehat{\theta},\theta) = \left\|\widehat{\theta}-\theta\right\|_2^2$ 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 1.1. *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 $v_1 \neq v_2$, $\rho(\theta(P_{v_1}, P_{v_2})) \geq 2\delta$. Let ℓ be as above. Then,*

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

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 $\overline{\mathbf{x}}$ as drawn from P_v^n.

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

$$T(\overline{\mathbf{x}}) \;:=\; \arg\min_{v\in\mathcal{V}}\; d\left(\widehat{\theta}(\overline{\mathbf{x}}),\theta(P_v)\right) \;.$$

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

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

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* $\widehat{\theta}$, it also we get that

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

which proves the claim. ■

2 Lower bounds via binary hyothesis testing (Le Cam's method)

We return to out 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.,

$$\widehat{\theta}(\overline{\mathbf{x}}) \;=\; \widehat{\theta}(x_1,\ldots,x_n) \;:=\; \frac{1}{n}\cdot\sum_{i\in[n]} x_i\,.$$

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

Exercise 2.1. *Let $P : \{0,1\} \to [0,1]$ be any distribution with $\mathbb{E}_{x\sim P}[x] = p(1) = \mu$. Show that*

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

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

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

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

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

3 Lower bounds for minimax rates via multiple hypotheses

We now consider a high-dimensional problem, where we can prove lower bounds using bounds for testing multiple hypotheses. Recall that for a random variable V uniformly distrbuted over a set of hypotheses $\mathcal{V}$, the probability of of error for any classifier $T(\overline{\mathbf{x}})$ with input $\overline{\mathbf{x}}$ coming from P_v^n for a randomly chosen $v \in \mathcal{V}$, is lower bounded as

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

As before, we will combine the above bound with Lemma 1.1 to prove the desired lower bound on the minimax rate using

$$\mathcal{M}_n(\Pi, \ell) \;=\; \inf_{\widehat{\theta}} \sup_{P \in \Pi} \mathop{\mathbb{E}}_{\overline{\mathbf{x}} \sim P^n}\left[\ell(\widehat{\theta}(\overline{\mathbf{x}}), \theta(P))\right] \;\geq\; \Phi(\delta) \cdot \inf_T \left\{\mathbb{P}\left[T(\overline{\mathbf{x}}) \neq V\right]\right\}$$

To use the above bounds, we need to come up with a set of distributions which are far in terms of the property θ (so that the second bound is large), but close on average in terms of KL-divergence (so that the first bound is large). This is also known as the *local Fano* method since we derived the first bound using Fano's inequality, and are applying it by using (a local bound on) KL-divergence for every pair of distributions P_{v_1}, P_{v_2} (recall that we used convexity of KL-divergence to reduce to the local setting). You can find other variants of this method in the notes by Duchi [Duc16].

3.1 Gaussian mean estimation

While binary hypothesis testing was used show a bound for estimating the mean of Bernoulli random variables, the multiple hypotheses setting is often useful in considering high-dimensional problems. We take Π to be the set of d-dimensional Gaussian distributions as below

$$\Pi \;=\; \left\{N(\mu, I_d) \;\mid\; \mu \in \mathbb{R}^d\right\}.$$

Let the property θ be the mean as before, and let $\ell(\widehat{\theta}, \theta) \;=\; \left\|\widehat{\theta} - \theta\right\|_2^2$. We first check the expected loss for the empirical mean estimator.

Proposition 3.1. *Let $\widehat{\theta}(x_1, \ldots, x_n) = \frac{1}{n} \cdot \sum_{i \in [n]} x_i$. Then, for any $\mu \in \mathbb{R}^d$, we have that*

$$\mathop{\mathbb{E}}_{\overline{\mathbf{x}} \sim (N(\mu, I_d))^n}\left[\left\|\frac{1}{n}\sum_{i \in [n]} X_i \;-\; \mu\right\|_2^2\right] \;=\; \frac{d}{n}.$$

Proof: The proof is similar to the case of Bernoulli random variables. By the (pairwise) independence of the n samples, we have that

$$\begin{aligned}
\mathop{\mathbb{E}}_{\overline{\mathbf{X}}\sim(N(\mu,I_d))^n}\left[\left\|\frac{1}{n}\sum_{i\in[n]} X_i \;-\; \mu\right\|_2^2\right] &= \frac{1}{n^2}\cdot\sum_{i\in[n]}\mathbb{E}\left[\|X_i-\mu\|_2^2\right] + \frac{1}{n^2}\cdot\sum_{i\neq j}\mathbb{E}\left[\langle X_i-\mu, X_j-\mu\rangle\right] \\
&= \frac{n}{n^2}\cdot\mathop{\mathbb{E}}_{X\sim N(\mu,I_d)}\left[\|X-\mu\|_2^2\right] \\
&= \frac{n}{n^2}\cdot d \;=\; \frac{d}{n}\,.
\end{aligned}$$

■

We will apply the local Fano method to prove the optimality of the above bound in terms of both d and n. We first need the following expression for KL-divergence of two normal distributions.

Exercise 3.2. *Prove (using the chain rule) that*

$$D(N(\mu_1, I_d)\;\|\;N(\mu_2, I_d)) \;=\; \frac{1}{2\ln 2}\cdot\|\mu_1-\mu_2\|_2^2\,.$$

Thus, we need to find a large collection of distributions, equivalent to finding a large collection of means, such that for any two $\mu_1 \neq \mu_2$, we have that $\|\mu_1-\mu_2\|$ is somewhat large (to lower bound the loss), but still $\|\mu_1-\mu_2\|$ is small on average (to upper bound the average KL-divergence). This is the content of the following lemma.

Lemma 3.3 (Packing lemma)**.** *There exists a collection of vectors $\mathcal{V}\subseteq\mathbb{R}^d$ such that $|\mathcal{V}|\geq 2^d$ and for all $v_1, v_2\in\mathcal{V}, v_1\neq v_2$, we have*

$$\frac{1}{2} \;\leq\; \|v_1-v_2\|_2 \;\leq\; 2\,.$$

It is actually quite easy to prove the packing lemma above (with slightly weaker parameters) but we will take a slightly longer route through covering and packing numbers to illustrate a general method. We first the lower bound, assuming the packing lemma. Let $\mathcal{V}$ be a collection as in Lemma 3.3. We consider the set of distributions

$$\{N(4\delta\cdot v, I_d) \;\mid\; v\in\mathcal{V}\}\,.$$

We have that for all $P, P'\in\Pi$, $\|\theta(P)-\theta(P')\|\geq 2\delta$. Also, since $\|v-v'\|\leq 2$ for any $v, v'\in\mathcal{V}$, we get that for any $P, P'\in\Pi$, the means are at distance at most 8δ. Hence,

$$D(P\|P') \;=\; \frac{1}{2\ln 2}\cdot\|\mu-\mu'\|_2^2 \;\leq\; \frac{1}{2\ln 2}\cdot(64\delta^2) \;=\; \frac{32\delta^2}{\ln 2}\,.$$

Applying the lower bound on minimax loss in terms of KL-divergences gives

$$\mathcal{M}_n(\Pi,\ell) \;\geq\; \delta^2\cdot\left(1-\frac{n\cdot(32\delta^2/\ln 2)+1}{\log|\mathcal{V}|}\right) \;\geq\; \delta^2\cdot\left(1-\frac{n\cdot(32\delta^2/\ln 2)+1}{d}\right)\,.$$

3.2 Covering and packing numbers

Definition 3.4. *Let S be a set of points with a metric $\rho(\cdot,\cdot)$. A collection of points $\mathcal{C} \subseteq S$ is called a δ-*covering *of S (with respect to the metric ρ) if*

$$\forall x \in S, \exists y \in \mathcal{C} \quad \rho(x,y) \;\leq\; \delta\,.$$

*A set of points $\mathcal{P}$ is called a δ-*packing *if*

$$\forall x, y \in \mathcal{P}, x \neq y \quad \rho(x,y) \;>\; \delta\,.$$

*The size of the minimal δ-covering, denoted as $N(\delta, S, \rho)$, is called the δ-*covering number *of S and the size of the maximal δ-packing is called the δ-*packing number*. The quantity $\log N(\delta, S, \rho)$ is also called the metric entropy of S.*

Remark 3.5. *As was pointed out after the lecture, the inequality in at least one of the two definitions above needs to be strict, for the remaining argument below to make sense. Typically, the inequality in the packing definition is taken to be strict.*

We will take the required collection in Lemma 3.3 to be a $(1/2)$-packing of the unit ball in $\mathbb{R}^d$ (under the Euclidean distance). We will show a lower bound on the size of this collection (the packing number) by using a relationship between the packing and covering numbers.

Exercise 3.6. *For any set S, metric ρ and $\delta > 0$, show that*

$$M(2\delta, S, \rho) \;\leq\; N(\delta, S, \rho) \;\leq\; M(\delta, S, \rho)\,.$$

(**Hint**: First prove that an optimal δ-packing must also be a δ-covering.)

Let $B_d(x,r)$ denote the ball in $\mathbb{R}^d$ of radius r (in the Euclidean distance) with its center at x. We know that $\mathsf{Vol}\left(B_d(x,r)\right) = c_d \cdot r^d$ for some constant $c_d \geq 0$. Note that if $\mathcal{C} \subseteq B_d(0,1)$ is a δ-covering of $B_d(0,1)$, then

$$B(0,1) \;\subseteq\; \bigcup_{x \in \mathcal{C}} B_d(x,\delta)\,.$$

Thus, we have

$$c_d \;=\; \mathsf{Vol}\left(B_d(0,1)\right) \;\leq\; \sum_{x \in \mathcal{C}} \mathsf{Vol}\left(B_d(x,\delta)\right) \;=\; N\left(\delta, B_d(0,1), \|\cdot\|_2\right) \cdot c_d \cdot \delta^d\,.$$

Combining with the previous relationship between covering and packing numbers, this gives

$$M\left(\delta, B_d(0,1), \|\cdot\|_2\right) \;\geq\; N\left(\delta, B_d(0,1), \|\cdot\|_2\right) \;\geq\; \frac{1}{\delta^d}\,.$$

Thus, there exists a $(1/2)$-packing of $B_d(0,1)$ of size at least 2^d. Note that for any v_1, v_2 in the packing

$$\frac{1}{2} \;<\; \|v_1 - v_2\| \;\leq\; \|v_1\| + \|v_2\| \;\leq\; 2\,,$$

which proves the packing lemma.

3.3 Another proof of (a weaker) packing lemma

One can also sample points on the Boolean cube $\{-1,1\}^d$ to prove a weaker version of the packing lemma, which also suffices for our application. For some applications, this additional structure in the set of points may be helpful.

Lemma 3.7. *There exists a collection of vectors $\mathcal{V} \subseteq \frac{1}{\sqrt{d}} \cdot \{-1,1\}^d$ such that $|\mathcal{V}| \geq 2^{d/20}$ and for all $v_1, v_2 \in \mathcal{V}, v_1 \neq v_2$, we have*

$$\frac{1}{2} \;\leq\; \|v_1 - v_2\|_2 \;\leq\; 2\,.$$

Remark 3.8. *The above bound is a crude one and is just proved as an illustration. A more sophisticated sampling argument can prove a lower bound of* $\exp(d/8)$ *on the size of the set $\mathcal{V}$.*

Proof: Sample $y_1, \ldots, y_N \in \{-1,1\}^d$ uniformly and independently at random, and take $\mathcal{V} = \left\{\frac{1}{\sqrt{d}} \cdot y_1, \ldots, \frac{1}{\sqrt{d}} \cdot y_N\right\}$. For a fixed pair v_i, v_j, we have

$$\|v_i - v_j\| \;\leq\; \frac{1}{2} \;\Leftrightarrow\; \langle v_i, v_j \rangle \;\geq\; \frac{7}{8} \;\Leftrightarrow\; \langle y_i, y_j \rangle \;\geq\; \frac{7d}{8}$$

Since $\langle y_i, y_j \rangle$ is a sum of d independent variables in $\{-1,1\}$, we have by Chernoff-Hoeffding bounds that

$$\mathbb{P}\left[\langle y_i, y_j \rangle \;\geq\; \frac{7d}{8}\right] \;\leq\; 2^{-\frac{d}{6}\cdot\left(\frac{7}{8}\right)^2} \;\leq\; 2^{-d/8}\,.$$

By a union bound

$$\mathbb{P}\left[\exists i, j \in [N]\, \langle y_i, y_j \rangle \;\geq\; \frac{7d}{8}\right] \;\leq\; N^2 \cdot 2^{-d/8}\,,$$

The probability above is $o(1)$ for $N \ll 2^{d/16}$, and thus, with high probability, we have for all $i \neq j$, $\|v_i - v_j\| \geq \frac{1}{2}$. The upper bound on the distance also holds since all vectors lie on the unit sphere. ■

References

[Duc16] John Duchi, *Lecture notes on Information Theory and Statistics*, 2016. 1, 4