Information and Coding Theory **Autumn 2017**

Lecture 9: October 25, 2017

Lecturer: Madhur Tulsiani

1 Lower bounds for minimax rates via multiple hypotheses

In this lecture, we extend the ideas from the previous lecture to develop lower bounds using lower 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

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

Recall that the above bound can be used to lower bound the minimax risk for a loss function $\ell(\widehat{\theta}, \theta) = \Phi \circ d(\widehat{\theta}, \theta)$ and a set of distributions Π. We proved in the last lecture that for a set of distributions $\{P_v\}_{v \in \mathcal{V}} \subseteq \Pi$ with the property that $\forall v_1 \neq v_2$, $d(P_{v_1}, P_{v_2}) \geq 2\delta$, we have

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

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 aevrage 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].

1.1 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

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

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 1.1 *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-divergenge of two normal distributions.

Exercise 1.2 *Prove (using the chain rule) that*

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

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). It is actually quite easy for the setting above, but we will take a slightly longer route through covering and packing numbers to illustrate a general method.

1.2 Covering and packing numbers

Definition 1.3 *Let S be a set of points with a metric $d(\cdot, \cdot)$. A collection of points $\mathcal{C} \subseteq S$ is called a δ-*covering *of S (with respect to the metric d) if*

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

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

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

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

We will take the required collection of means to be a scaled copy of 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 relatioship between the packing and covering numbers.

Exercise 1.4 *For any set S, metric d and $\delta > 0$, show that*

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

(**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. We will use this to prove the lower bound for mean estimation.

1.3 Back to lower bounds for mean estimation

We can now get back to the lowe bound. Let $\mathcal{V}$ be a $(1/2)$-packing of $B_d(0,1)$ of size at least 2^d. We consider the set of distributions

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

Since $\mathcal{V}$ is a $(1/2)$-packing, 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,

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

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

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

1.4 Sparse mean estimation

From the previous examples, it seems like the empirical mean estimator is always the best one, and the role of information theory is primarily for proving lower bounds. However, it can also serve as a guide for the right bound to aim for. Consider the set of normal distributions, where the mean has only *one* non-zero coordinate.

$$\Pi \;=\; \left\{N(\mu, I_d) \;\middle|\; \mu \in \mathbb{R}^d,\; \|\mu\|_0 \leq 1\right\}.$$

Let $\theta(P) = \mathbb{E}_{x\sim P}[x]$ be the mean, and let $\ell(\widehat{\theta}, \theta) = \left\|\widehat{\theta} - \theta\right\|_2^2$ as before.

Exercise 1.5 *Let $\mathcal{V} = \{e_1, \ldots, e_d\}$ be the set of standard basis vectors in $\mathbb{R}^d$. Use the set of means $\mu_v = \sqrt{2} \cdot v$ for $v \in \mathcal{V}$ to show that there exists a constant c such that*

$$\mathcal{M}_n(\Pi, \ell) \;\geq\; c \cdot \frac{\log d}{n}\,.$$

The optimal estimator for the above problem actually extends the definition of the mean as the minimizer of the total square distance (from the sample points). For a sample $\overline{\mathbf{X}} \in (\mathbb{R}^d)^n$ and $j \in [d]$, we define $\widehat{\mu}^{(j)}$ as a vector which is equal to the empirical mean in the j^{th} coordinate and zero elsewhere i.e.,

$$\widehat{\mu}_k^{(j)} = \begin{cases} \frac{1}{n} \cdot \sum_{i\in[n]} (X_i)_j & \text{if } k = j \\ 0 & \text{otherwise} \end{cases}$$

We take the estimator to be

$$\widehat{\mu} \;=\; \operatorname*{arg\,min}_{\mu \in \left\{0, \widehat{\mu}^{(1)}, \ldots, \widehat{\mu}^{(d)}\right\}} \left\{\frac{1}{n} \cdot \sum_{i\in[n]} \|X_i - \mu\|_2^2\right\}.$$

This estimator indeed achieves an expected loss of $O((\log d)/n)$, but we will not discuss the proof here.

References

[Duc16] John Duchi. Lecture notes on Information Theory and Statistics, 2016. URL: https://stanford.edu/class/stats311/. 1